20世纪统计怎样变革了科学



女士品茶

David Salsburg 萨尔斯伯格（美）著

邱东 等译



女 士 品 茶

The Lady Tasting Tea

How Statistics Revolutionized Science in the Twentieth Century

20 世纪统计怎样变革了科学

David Salsburg 萨尔斯伯格（美）著

邱东 等译

目　录

作者序
第 1 章　女士品茶
第 2 章　偏斜分布
第 3 章　可爱的戈塞特先生
第 4 章　在“垃圾堆”中寻觅
第 5 章　收成变动研究
第 6 章　“百年不遇的洪水”
第 7 章　费歇尔获胜
第 8 章　致死的剂量
第 9 章　钟形曲线
第 10 章　拟合优度检验
第 11 章　假设检验
第 12 章　置信诡计
第 13 章　贝叶斯异论
第 14 章　数学界的莫扎特
第 15 章　“小人物”之见解
第 16 章　非参数方法
第 17 章　当部分优于总体时
第 18 章　吸烟会致癌吗？
第 19 章　如果您需要最佳人选……
第 20 章　朴实的德克萨斯农家小伙
第 21 章　家庭中的天才
第 22 章　统计学界的毕加索
第 23 章　处理有瑕疵的数据
第 24 章　重塑产业的人
第 25 章　来自黑衣女士的忠告
第 26 章　鞅的发展
第 27 章　意向治疗法
第 28 章　电脑随心所欲
第 29 章　“泥菩萨”

附：作者后记
大事年表

Table of Contents

Chapter 01　The Lady Tasting Tea

Chapter 02　The Skew Distribution

Chapter 03　That Dear Mr. Gosset

Chapter 04　Raking Over the Muck Heap

Chapter 05　“Studies in Crop Variation”

Chapter 06　“The Hundred-Year Flood”

Chapter 07　Fisher Triumphant

Chapter 08　The Dose That Kills

Chapter 09　The Bell-Shaped Curve

Chapter 10　Testing the Goodness of Fit

Chapter 11　Hypothesis Testing

Chapter 12　The Confidence Trick

Chapter 13　The Bayesian Heresy

Chapter 14　The Mozart of Mathematics

Chapter 15　The Worm’s-Eye View

Chapter 16　Doing Away With Parameters

Chapter 17　When Part is Better than the Whole

Chapter 18　Does Smoking Cause Cancer

Chapter 19　If You Want the Best Person

Chapter 20　Just A Plain Texas Farm Boy

Chapter 21　A Genius in the Family

Chapter 22　The Pieasso of Statistics

Chapter 23　Dealing with Contamination

Chapter 24　The Man Who Remade Industry

Chapter 25　Advice From the Lady in Black

Chapter 26　The March of the Martingales

Chapter 27　The Intent to Treat

Chapter 28　The Computer Turns Upon Itself

Chapter 29　The Idol With Feet of Clay

作者序

进入 19 世纪时，科学界奉行着一种固化的哲学观，即机械式宇宙观（clockwork universe）。这种哲学观认为，为数不多的几个数学公式，像牛顿的运动定律（Newton's laws of motion）和玻意耳的气体定律（Boyle's laws of gases），可以用来描述现实世界的一切，并能预测未来即将发生的事件。而对这种预测，所需要的不过是一套完整的公式，以及一组具有足够精确度的相关数据。然而，对于一般大众来说，整整花了 40 年时间，他们的思想才跟上这种科学观念。

这种思想上的落差，典型地体现在 19 世纪早年拿破仑皇帝（Emperor Napoléon）与皮埃尔·西蒙·拉普拉斯（Pierre Simon Laplace）的一次对话中。拉普拉斯写了一本历史性的权威著作，论述如何根据地球上少数观察数据来计算行星和彗星的未来位置。据说拿破仑问道："拉普拉斯先生，我发现你的论述中没有提到上帝啊！"拉普拉斯的回答则是："我不需要这个假设条件。"

机械式宇宙观认为，宇宙如同一个庞大的时钟机器，所有的物体都按照一定的规律运动，宇宙永续运转而不需要神的介入；所有将来发生的事件都决定于过去的事件。许多人对这种无神论的思想感到恐慌，从某种意义上说，19 世纪浪漫主义运动的兴起，正是对这种精确应用推理的冷冰冰的哲学观的回应。然而，19 世纪 40 年代出现了对新科学的证明，这叫一般人难以想象：牛顿的数学定律被用来预测另一颗行星的存在，而海王星（the planet Neptune）正是在这些定律所预测的位置被发现的。于是，几乎所有对机械宇宙观的反抗都被粉碎了，这一哲学立场很快成为大众文化的基本部分。

不过，就算拉普拉斯在他的公式中不需要上帝，他还是需要一种被他称为误差函数（error function）的东西。从地球上对行星和彗星的观察，与用公式所预测的位置并不绝对吻合，拉普拉斯和他的科学家同伴将这归结于观察中的误差，有时是由于地球大气层中的扰动，有时则是人为的。拉普拉斯把所有这些误差都放在一个附加项（误差函数）里，从而将之纳入他的数据描述。这个误差函数吸收了所有的误差，剩下的只是用来预测宇宙星体实际位置的绝对运动定律。当时科学家相信，随着越来越精确的测试，对误差函数的需求将逐渐消失。由于有误差函数来表示预测值与观察值之间的微小差异，19 世纪早期的科学可以说是受到了哲学上决定论（determinism）的掌控，即相信所发生的任何事情都预先地决定于两点：（1）宇宙的初始条件；（2）描绘其运动的数学公式。

到了 19 世纪末，误差并没有消失，反倒是增加了。当测试越来越精确，误差也越来越多。机械宇宙观处于动摇之中，试图发现生物学定律和社会学定律的努力也失败了。在物理和化学等传统科学中，牛顿和拉普拉斯所用的那些定律，逐渐地被证明只是粗略的逼近。这样，科学便渐渐开始在新的范式（paradigm）下运作，这新范式就是现实世界的统计模型。到 20 世纪末期，几乎所有科学都转而运用统计模型了。

大众文化还是没有跟上这种科学革命，尽管一些含混的观念和表述，像相关（correlation）、胜率（odds）和风险（risk）等等，已经渗入了大众的词汇，并且多数人意识到了不确定性问题，这是与诸如医学和经济学等学科领域相联系的。但就已经发生的哲学观的深层转变而言，学界之外没有人能够对此有什么理解。这些统计模型是什么？它们是怎么来的？在现实生活中它们意味着什么？它们是现实的真实描述吗？本书正是试图来回答这些问题，其中我们也想介绍一些先生和女士的生平故事，这些人曾涉身于这场革命之中。

在处理这些问题时，必须把三个数学概念区分开：随机（randomness）、概率（probability）和统计（statistics）。对大多数人而言，随机只是不可预测性（unpredictability）的另一个说法。犹太教法典（Talmud）中的一则格言，传达了这种通

常的看法：“不应该去探寻宝藏，因为宝藏的发现是随机的；按照定义，没有人能够寻找只会被随机发现的东西。”但是，对现代科学家来说，随机性有许多不同的类型。概率分布（probability distribution，这将在第 2 章中讨论）的概念允许我们对随机性加以限制，并赋予我们有限的能力去预测未来的随机事件。因此，对现代科学家而言，随机事件并不是杂乱的、不可预期的和不可预测的，它们有一个可以用数学来描述的结构。

概率是一个非常古老概念的现代用语，它曾出现在亚里士多德（Aristotle）的著作中。这位先哲声称：“不可能事件将会发生，这正是概率的特性。”起初，概率只是涉及到个人对什么事件即将发生的预测，在 17 和 18 世纪，一批数学家，其中包括贝努里（Bernoullis）父子、费尔马（Fermat）、棣莫弗（de Moivre）、帕斯卡（Pascal）都在以机会博弈（games of chance）为起点去研究概率的数学理论。他们发明一些非常高级的方法，用来计算等可能事件，棣莫弗设法在这些技术中加进微积分的方法，贝努里则可以领悟出非常基础的定理，叫大数定律（Laws of large numbers）。到了 19 世纪末期，数理概率主要由一些非常高级的技巧构成，但还缺少坚实的理论基础。

尽管不够完善，还是可以证明概率理论对发展统计分布（statistics distribution）观念的作用。当我们考虑一个特殊的科学问题时，就会产生一个统计分布。例如，在 1971 年，哈佛公共卫生学院所做的一项研究发表在英国的医学期刊《柳叶刀》（Lancet）上，这项研究旨在检验喝咖啡是否与下泌尿道癌有关。研究的报告以一级病人为对象。其中一些人患有下泌尿道癌，另一些人则患有其它疾病。报告的作者还搜集了这组病人的其它资料，如年龄、性别和家族的癌症病史等。结果证明，并不是每个喝咖啡的人都会得泌尿道癌，也不是每个得泌尿道癌的人都圆角咖啡，所以存在着与他们的假设相矛盾的事件。然而，25%的此类癌症患者习惯每天喝 4 杯以上咖啡，只有 10%的非癌症患者是这种咖啡嗜好者，因而，似乎有一些证据支持这种假设。

这种资料的搜集给研究者提供了一个统计的分布。运用数理概率的工具，他们为这个分布建造了一个理论公式，称之为概率分布函数（probability distribution function），或简称分布函数（distribution function），以此来检验所研究的问题。它与拉普拉斯的误差函数相似，但却复杂许多。运用概率论来建造理论分布函数，而这个函数用来描述从未来数据中所能得到的预期结果，这些数据是以随机方式从同一总体的人群中提取的。

我不想使本书成为一本关于概率和概率论的书，那是抽象的数据概念。本书涉及的一些概率定理在科学问题上的应用，涉及统计分布和分布函数的世界。概率论本身不足以说明统计方法，有时甚至会出现这样的情形：科学中所用的统计方法违背了概率的定理。读者会发现本书中概率时隐时现，需要时被用到，不需要时则被忽略。

由于现实世界的统计模型都是数学化的，充分理解它们只能用数学公式或符号的方式。本书是一种野心不那么大的尝试，我打算描述发生在 20 世纪科学界的统计革命，而手法是通过介绍一些参加过这场革命的人物（其中不少人至今还健在）。我只是涉猎他们创造性的工作，试图让读者从中体会他们的个别发现是如何适应整个统计革命的。

仅就本书而言，读者并不会学到对科学数据进行统计分析所需要的足够知识，那需要几年的循序渐进的学习。但我希望读者看过本书后，能够对科学的统计观所代表的基本哲学的重大变革有所理解。那么，不懂数学的人要理解这场科学革命，应该从哪里开始呢？我以为，一个不错的选择是与女士一道品茶。

第 1 章 女士品茶

那是 20 世纪 20 年代后期，在英国剑桥一个夏日的午后，一群大学的绅士和他们的夫人们，还有来访者，正围坐在户外的桌旁，享用着下午茶。在品茶过程中，一位女士坚称：把茶加进奶里，或把奶加进茶里，不同的做法，会使茶的味道品起来不同。在场的一帮科学精英们，对这位女士的“胡言乱语”嗤之以鼻。这怎么可能呢？他们不能想象，仅仅因为加茶加奶的先后顺序不同，茶就会发生不同的化学反应。然而，在座的一个身材矮小、戴着厚眼镜、下巴上蓄着的短尖髯开始变灰的先生，却不这么看，他对这个问题很感兴趣。

他兴奋地说道：“让我们来检验这个命题吧！”并开始策划一个实验。在实验中，坚持茶有不同味道的那位女士被奉上一连串的已经调制好的茶，其中，有的是先加茶后加奶制成的，有的则是先加奶后加茶制成的。

写到这里，我可以想象，部分读者会对这种实验不以为意，认为它不过是一帮精英们于夏日午后的一个小消遣。他们会说：“这位夫人能不能区分两种不同的注茶方式，又有什么大不了的呢？这个问题并没有什么科学价值，这些大人物更应该把他们的天才用在对人类有所裨益的事情上去。”

不幸的是，不管外行对科学及其重要性怎么想象，从我个人的经验来看，大多数科学家之所以从事科研活动，只是因为他们对结果感兴趣，或者能够在工作中得到理性的刺激。好的科学家很少会想到工作的最终重要性，剑桥那个晴朗夏日的午后也是这种情景。那位夫人也许能、也许不能正确地品出不同的茶来，但这无关紧要，因为，实验的真正乐趣，在于找到一种判断该女士是对还是错的方案来。于是，在蓄着胡须先生的指导下，大家开始讨论应该如何进行实验判断。

接下来，在场的许多人都热心地加入到实验中来。几分钟内，他们在那位女士看不见的地方调制出不同类型的茶来。最后，在决战来临的气氛中，蓄短胡须的先生为那位先生为那位女士奉上第一杯茶，女士品了一小会儿，然后断言这一杯是先倒的茶后加的奶。 这位先生不加评论地记下了女士的说法，然后，又奉上了第二杯……

科学的合作性质

这个故事是我在 20 世纪 60 年代后期，从一个当时在场的先生那里听到的。这位先生就是休・史密斯（Hugh Smith），但他都是以 H・费尔菲尔德・史密斯（H. Fairfield Smith）的名义发表科研论文。我认识他的时候，他在位于斯托尔斯（Storrs）的康涅狄格大学（the University of Connecticut）任统计学教授，而我则是两年以前在这个大学拿到了统计学博士学位。在宾州大学（the University of Pennsylvania）教了一阵子书后，我加入到了辉瑞公司（Pfizer Inc.）的临床研究部门。这是一家大型制药公司，它的研究园区坐落在格罗顿（Groton），离斯托尔斯大约一个小时的车程。当时，我是那里唯一的统计学家。在辉瑞期间，我要处理许多疑难的数学问题，还要负责给他们讲解这些问题，并告诉他们，对这些问题，我个人的结论是什么。

在辉瑞工作期间，我发现，科研工作几乎不能独立完成，通常需要不同智慧的结合。因为，这些研究太容易犯错误了。当我提出一个数学公式作为解决问题的工具时，这个模型有时可能并不适合；或者我就所处理情况而引入的假设并不真实；或者我发现的“解”是公式中的失误部分推导出来的；甚至我可能在演算中出了错。

无论何时，我去斯托尔斯的大学拜访，与史密斯教授探讨问题，或者，与辉瑞的化学专家、药理专家坐在一起讨论，我提出的问题都会受到欢迎，他们对这种讨论充满兴趣和热情。对大多数科学家来说，工作中令他们最感兴趣的，就是解决问题时那种兴奋感。因此，

在检验并试图理解问题时，他们期盼着与他人交流。

实验的设计

剑桥那个夏日午后的情形正是如此，那个留着短胡须的先生就是罗纳德·艾尔默·费歇尔（Ronald Aylmer Fisher），当时他只有三四十岁。后来，他被授予爵士头衔。1935 年，他写了一本叫《实验设计》（The Design of Experiments）的书，书的第 2 章就描述了他的“女士品茶”实验。在书中，他把女士的断言视为假设问题，他考虑了各种可能的实验方法，以确定那位女士是否能做出区分。设计实验时的问题是，如果只给那位女士一杯茶，那么即使她没有区分能力，她也有 50%的机会猜对。如果给两杯茶，她仍可能猜对。事实上，如果她知道两杯茶分别以不同的方式调制，她可能一下子全部猜对（或全部猜错）。

同样，即便这位女士能做出区分，她仍然有猜错的可能。或者是其中的一杯与奶没有充分地混合，或者是泡制时茶水不够热。即便这位女士能做出区分，也很有可能是奉上了 10 杯茶，她却只是猜对了其中的 9 杯。

在这本书中，费歇尔讨论了这个实验的各种可能结果，他叙述了如何确定这样一些问题：应该为那位女士奉上多少杯茶？这些茶应该按什么样的顺序奉上？对所奉各杯茶的顺序应该告诉那位女士多少信息？依据那位女士判断的对错与否，费歇尔搞出了各种不同结果的概率。但在讨论中，他并没有指明这种实验是否真的发生过，也没有叙述这次实验的结果。

费歇尔书中有关实验设计的著述是科学革命的要素之一，这场革命在 20 世纪前半叶席卷了科学的所有领域。早在费歇尔出道以前，科学实验已经进行了几百年。在 16 世纪后期，英国的威廉•哈维（William Harvey）用动物做实验，他将不同动物静脉和动脉里的血液堵住，试图追踪血液从心脏到肺，回流到心脏，流向全身，再回到心脏的循环路线。

费歇尔没有发现实验是增长知识的方法。费歇尔之前，实验对每个科学家而言都是有其特性的。优秀的科学家可以做出产生新知识的实验，而二流的科学家常常从事的是积累数据的实验，但对知识增长没有什么用处。为说明这点，可以举发生在 19 世纪后期的一个例子。那时的科学家就测量光速做了许多无关要旨的努力，而直接到美国物理学家艾伯特·米切尔森（Albert Michelson）用光线和镜子建造了一个特别精巧的系列实验，才第一次得到好的估计。

在 19 世纪，科学家很少发表实验结果。他们所做的是论述自己的结论，并发表能证明结论真实性的数据。格雷戈尔•门德尔（Gregor Mendel）没有展示出他全部豌豆培育实验的结果，他叙述了他的系列实验，然后写道：“两组系列实验的前 10 个数据可以用来说明……”在 20 世纪 40 年代，费歇尔检验了门德尔用来说明结论的数据，发现这些数据过分完美，以至于失真，它们并没有表现出应该具有的随机程度。

尽管科学从审慎思考、观察和实验发展而来，但从来不清楚应该怎样从事实验，实验的全部结果通常也没有展现给读者。

19 世纪末和 20 世纪初的农业研究中，上述情况尤为明显。20 世纪早期费歇尔在农业实验站工作，在费歇尔去那儿工作之前，这个实验站已经进行了约 90 年的肥料构成（称之为人工肥料）实验。在一个典型的实验中，工人将磷肥和氮肥的混合物撒在整块田中，然后种植作物，测度收成和整个夏季的雨量。这里有精巧的公式用来“调整”某年或某块地的产量，以便与另一块地、或同一块地的另一年产量相比，这被称为“肥力指数”。每一个农业实验站都有自己的肥力指数，而且都认为自己的指数是最精确的。

90 年的实验结果不过是一堆未经发表、了无用处的混乱数据。看来某些品种的小麦对某种肥料反应优于其它品种，但只是在降雨过量的年份如此。其它实验似乎显示：第一年用钾硫化物，第二年用碳酸硫化物，会使某些品种的马铃薯增产，而对其它品种并非如此。因

此，就这些人工肥料，充其量可以说，其中有些在有的时候，可能或大概有效。

作为一个卓越的数学家，费歇尔审视了农业科学家用来修正实验结果的肥力指数，这些指数是用来解释不同年份气象变化所造成的差异的，他还检查了其它农业实验站所用的同类指数。当简化为基本的代数式时，这些指数不过是同一公式的不同表现形式，换句话说，看似激烈争斗的两个指数，其实起着同样的修正作用。1921 年，费歇尔在农业科学领域的领军期刊《应用生物学年报》(the Annals of Applied Biology)上发表了一篇论文，文中他指出了采用哪种指数并没有什么差异，并且，所有修正都不足以调整不同地块上的肥力差异。这篇非凡的论文终止了一场持续 20 多年的科学论战。

费歇尔接着检查了过去 90 年来的雨量和收成数据，指出年度间不同气候的影响远远大于不同肥力的影响。用费歇尔后来在他的实验设计理论里发明的一个词来说，“混合”（confounded）的，这意味着用已有的实验数据是不能将二者分开的。90 年的实验和 20 年的科学论战几乎是无谓的浪费。

这使得费歇尔专注于实验和实验设计的思考。他的结论是：科学家需要从潜在实验结果的数据模型开始工作，这是一系列数据公式，其中一些符号代表实验中将被搜集的数据，其它则代表实验的全部结果。科学家从实验数据开始，并计算与所考虑科学问题相应的结果。

让我们考虑一个关于一个老师和某个学生的简单例子。这个老师非常想找出一些关于这个孩子学习情况的测试数据，为了达到这个目的，老师对孩子进行了一组考试，每一个考试都在 0 到 100 之间评分，任何一个单一的考试都不可能对孩子知识的掌握提供可靠的评估；这个孩子可能是没有学习多少考试所涉及的内容，但是知道不少考试以外的事情；可能是这个孩子在参加考试那天头疼；还可能是参加考试那天早上孩子与父母发生了争执。由于种种原因，单一考试不能对知识量提供好的估计，所以老师进行了一组考试，然后计算出所有考试的平均分来评价孩子的知识量。这样的估计结果会更好，多少分是孩子知识量的实验结果，而每一个单独考试的分数则是数据。

那么老师应该如何组织考试？是搞那种只包括几天前所教授内容的系列考试，还是每次考试都从考试前所教授的全部内容中提取一部分？考试是一个星期搞一次，还是每天搞一次？或者在每个教学单元结束时搞？所有这些都是实验设计涉及到的问题。

如果农业科学家想知道某种人工肥料对小麦生长的效用，就要构建一个实验以取得效用估计时所需要的数据。费歇尔表明，实验设计的第一步是建立一组数学公式，用以描述待搜集数据与欲估计结果之间的关系，因此，任何有用的实验必须是能够提供估计结果的。实验必须是有效的，能够让科学家测定出气候的差异和不同肥料的使用对产量差别的影响。特别是，有必要包括同一实验中打算加以比较的实验处理（treatments）,即那些后来被称为“控制组件”（controls）的东西。

在他那本关于实验设计的书中，费歇尔提供了几个实验设计的范例，并导出优秀设计的一般原则。然而，费氏方法中所涉及到的数学非常复杂，多数科学家设计不了自己的实验，除非他们遵循费歇尔书中提出的实验设计中的某个模式。

农业科学家认识到费歇尔工作的伟大价值，在大多数说英语的国家中，费氏方法很快便成为农业科研的主流学派。从费歇尔的原创性工作出发，用来论述不同实验设计的完整科学文献发展起来。这些设计被应用到农业以外的领域，包括医学、化学和工业质量管理。在许多案例中，所涉及的数学高深且复杂，但此时此刻，我们不妨停下来想想，科学家不可能不假思索地动手实验，这通常需要长时间的审慎思考，而且，其中通常会有大量的、高难的数学。

至于前面所说的女士品茶——那个在剑桥晴朗的夏日午后所做的实验中，那位女士怎样了呢？费歇尔没有描述这项实验的结果，但史密斯教授告诉我，那位女士竟然正确地分辨出了每一杯茶！

第 2 章 偏斜分布

像人类思想史上的许多革命一样，要想找到统计模型成为科学组成部分的确切时刻，也是很难的。人们可以在 19 世纪初德国和法国数学家的工作中找到可能存在的特例，甚至在 17 世纪伟大的天文学家约翰尼斯·开普勒（Johannes Kepler）的论文中，也能找到某种启示。正像本书前言中所提到的那样，拉普拉斯（Laplace）发明了误差函数来说明天文学中的统计问题，但我仍然倾向于把统计革命的发生定位于 19 世纪 90 年代 K·皮尔逊（Karl Pearson）的工作。查尔斯·达尔文（Charles Darwin）把生物变异认作生命的基本面，并将之作为适者生存理论的基础。然而，是他的英国伙伴 K·皮尔逊首先认识到统计模型的根本性质，以及这种模型对 19 世纪科学中的决定论观点提供了哪些不同的东西。

当我在 20 世纪 60 年代开始学习数理统计时，K·皮尔逊的名字在课上很少被提到。当我与这一领域的大人物共同探讨一些问题时，也听不到对 K·皮尔逊及其著作的参考。他或者是被忽略了，或是被视为行为早已出局的次要人物。例如，美国国家标准局（the U.S. National Burean of Standards）的邱吉尔·艾森哈特（Churchill Eisenhart）当时正在伦敦大学学院（University College,London）学习，那是 K·皮尔逊人生的最后几年，艾森哈特记忆中的 K·皮尔逊不过是一个精神头不足的老头儿。统计研究的步伐已经将他推出局外，他和他的工作被埋进故纸堆中，青年学生神采飞扬，集聚在新的大人物周围学步，其中之一，便是 K·皮尔逊自己的儿子，但是没有人去拜见老皮尔逊，他的办公室孤零零地坐落在那里，远离着活跃的、振奋人心的新研究。

当然并不总是如此，在 19 世纪 70 年代，年轻的 K·皮尔逊离开英国，到德去从事政治科学的研究生学习。在那里，他倾心于卡尔·马克思（Karl Marx）的著作，为了表达崇拜之情，他把自己名字的拼法从 Carl 改成 Karl。带着政治学博士的学位，他回到了伦敦，并在这个领域写过两本值得重视的著作。在维多利亚时代的英国，伦敦的拘谨之风最甚，K·皮尔逊却大胆地效仿德国和法国上流社会的沙龙，组织了一个青年男女谈话俱乐部（Young Mens and Womens Discussion Club）。俱乐部的青年男女平等地聚焦在一起（未婚少女并没有人陪伴），讨论世界上重大的政治和哲学问题。K·皮尔逊正是在那种环境下与夫人相遇而结缘的，这个事实使人感到发起这类俱乐部可能另有动机。这个小小的社会冒险对我们进入 K·皮尔逊的内心世界提供了帮助，可以见证他对已经建立起来的传统是那样地不以为意。

尽管拿的是政治学博士学位，K·皮尔逊的主要兴趣还是在科学哲学和数学模型的性质上。19 世纪 80 年代，他发表了《科学的法则》（The Grammar of Science），这本书后来再版了多次。在第一次世界大战之前的一段时间里，它被视为关于科学和数学性质最伟大的著作之一，其中充满了闪光的、原创性的、最具洞察力的见解，这使该书成为科学哲学的一本重要著作。同时，它又是以流畅、简单的风格写成，任何人都可以接受，你不必懂得数学就可以理解《科学的法则》。尽管从写作之日算起，这本书已经有 100 多年的历史了，但其中充满洞察力的见解和思想，对 21 世纪的数学研究，仍然是适用的。而它所提供的对科学性质的理解，至今也是真实的。

高尔顿的生物统计实验室

在人生的这个时段，K·皮尔逊感受到了英国科学家弗朗西斯·高尔顿（Francis Galton）爵士的影响。大多数人知道高尔顿这个名字，缘于他是指纹现象的“发现者”。高尔顿的贡献是认识到指纹对每一个人都是独特的，此外，还有通常用于识别和分类指纹的方法。指纹的唯一性存在于手指类型中出现的不规则标识和切面，这被称为“高尔顿标识”（Galton Marks）。高尔顿做的远比这多，作为一个只是将生物学算作其业余爱好的科学家，通过数字

模型的研究，他寻求将数学的严密引入生物学，这同样是富有价值的。他所初创的各种调查当中的一项，是对天才遗传的研究。在这项研究中，他搜集了有关父子的信息，这些人因智商高而闻名。但由于当时对智力的测量没有什么好的办法，他发现研究这个问题特别困难，于是他决定转向诸如身高之类的遗传特性的研究，因为这更容易测量些。

高尔顿在伦敦成立了生物统计实验室（biometrical laboratory），并打广告动员不同的家庭来做测量。在这个实验室，他搜集身高、体重数据，测量特殊的骨骼和家庭成员的其它特性。他和他的助手将这些数据列成表格，并一再检验，他是在寻找利用父母测度数据来推断子女的某些办法。比如说，很明显，高个子父母很容易有高个子的小孩，但是不是存在某些数学公式，只用父母的身高就可以预测孩子将有多高呢？

相关与回归

高尔顿用这种方法，发现了他称之为“向平均回归”（regression to the mean）的现象，这表现为：非常高的父亲，其儿子往往要比父亲矮一些；而非常矮的父亲，其儿子往往要比父亲高一些。似乎是某种神秘的力量，使得人类的身高从高矮两极移向所有人的平均值。不只是人类身高存在着向平均数回归的现象，几乎所有的科学观察都着了魔似的向平均值回归。在第 5 章到第 7 章，我们将看到，费歇尔如何能够将高尔顿向平均值回归的思想纳入统计模型，而这种模型现在支配着经济学、医学研究和工程学的很多内容。高尔顿仔细思考了他的惊人发现，而后认识到这必定是真实的，在进行所有观察之前这就是可以预言的。他说，假设不发生这种向平均值的回归，那么从平均意义上看，高身材父亲的儿子将与他们的父亲一样高，在这种情况下，一些儿子的身材必须高于他们的父亲，以抵消身材比父亲矮小者的影响，使平均值不变。高身材者这一代人的儿子也将如此，那么会有一些儿子身材更高。这个过程将一代一代延续下去。同样地，将会有一部分儿子身材比他们的父亲矮小，而且有一部分孙子将更加矮小，如此下去，不用多少代，人类种族就将由特别高和特别矮的两极构成。

上述的情形并没有发生，人类的身高在平均意义上趋向于保持稳定。只有当非常高的父亲其儿子平均身材变矮，而非常矮的父亲其儿子的平均身材变高，才能出现这种稳定。向平均值回归是一种保持稳定性的现象，它使得某给定物种代际之间大致相同。

高尔顿发现了这种关系的一种数学测度，他称之为“相关系数”（coefficient of correlation）。高尔顿给出了明确的公式，以计算这个系数，所用的资料则是在生物测量实验室搜集的。这是一个非常详细而明确的公式，它只计算了向平均值回归的一个方面，但没有告诉我们任何有关这种现象原因的信息。正是在这个意义上，高尔顿最先使用了“相关”这个字眼，这之后它演变进入了大众词汇。与高尔顿特定的相关系数相比，“相关”经常被用来表示更为模糊的东西，尽管“相关”本身有严格的科学含义。科学圈外的人经常说到这个词，似乎它描述了两种事物如何相联系，但除非你涉及到高尔顿的数学测量，否则，当你使用高尔顿用于特别目的的“相关”这个词时，它不必那么精确。

分布与参数

有了这个计算相关的公式，高尔顿实际上已经非常接近新的革命性观念了，这个观念革命在 20 世纪几乎修正了所有的学科。但却是他的弟子 K·皮尔逊，在非常完整的意义上第一个规范地阐明了这个观念。

为了理解这个革命性的观念，你必须将已有的关于科学的成见抛开。通常我们被教导，科学就是测量，我们进行精心的测量，并用它来寻找描述自然的数学公式。在高中的物理课中我们学过，当时间给定时，一个自由落体的运行将遵循一个含有符号“g”的公式，这里

的“g”是关于重力加速度的常量。我们学过可以用来确定“g”的值的实验。然而，当高中生们进行一系列确定值的实验时，顺着斜板滚动小球，并测量小球需要多长时间到达不同的位置时，发生了什么呢？这就是很少得出确切的结果。学生进行实验的时间越长，困惑就越多，因为不同的实验得出了不同的“g”值。老师仅凭自己优越的知识来审视学生的实验，并认定学生之所以得不到正确的结果，要么是因为工作草率，要么是因为不够细致，要么是抄错了数据。

老师没有告诉学生的是：所有的实验都是草率的，并且，即使是最精心的科学家，也很少得到确切的数值。不可预见和不可观察的小扰动在每一个实验中都有：室内的空气可能太潮湿，或者落体在滚动前卡住了一个微秒，旁边飞过的蝴蝶可能会有其影响：造成气流的轻微扰动。人们从一个实验中真正得到的是散乱的数据，其中没有一个单个数据是确切的，但所有这些数据可以用来对确切值进行近似的估计。

武装了 K·皮尔逊的革命性观念，我们就不再将实验结果看作精心测量得出的数据，它们也不是本来就确切的，用更容易接受的术语来代替：它们是一组散布数据，或一个数据分布中的样本。数据的分布可以写成数学公式，它告诉我的数值是不可预测的，我们只能谈论概率值而不是确定值，单个实验的结果是随机的，在这个意义上看它们是不可预测的，然而，分布的统计模型却使我们能够描述这种随机的数学性质。

科学家花了一些时间才认识到观测值所固有的随机性质。在 18 和 19 世纪，天文学家和物理学家创造出描述他们观察值的数学公式，达到了可接受的精确程度，在为测量工具不够精确，所以观察值与预测值之间的是预料之中的，可以忽略不计。星体和其它天体的运动被假定遵循运动基本公式所确定的精确路径，其不确定性是由于简陋的测量工具造成的，并不是其固有的性质。

随着物理学中更为精确的测量工具的发展，随着将这种测量科学扩展到生物学和社会学的尝试，大自然所固有的随机性越来越明显了。怎么处理它？一种办法是坚持数学公式的精确性，将观测值与预测值之间的离差视为小的、无关紧要的误差。事实上，早在 1820 年，拉普拉斯的数学论文描述了第一个概率分布，即误差分布，那是一个与这些小的、无关紧要的误差相联系的概率的数学公式。这个误差分布以钟形曲线（bell-shaped curve）或正态分布（the normal distribution）[1]的说法进入了大众的词汇。

这使 K·皮尔逊比正态分布或误差分布更进了一步，审视生物学中积累的数据。K·皮尔逊认为，测量值本身，而不是测量的误差，就具有一种正态分布。我们所测量的，实际上是随机散布的一部分，它们的概率通过数学函数——分布函数被描述出来。K·皮尔逊发现了被他称为“偏斜分布”（skew distribution）的一组分布函数，他宣称，这组函数可以描述科学家在数据中可能遇到的任何散布类型，这组函数中的每一个分布由四个数字所确定。

用来确定分布函数的这些数字与测量中的数字不属于同一类型，这些数字决不会被观察到的，但可以从观测值散布的方式中推导出来。这些数字后来被称为参数（parameters——源自希腊语，意思是“几乎测量”（almost measurements））。能够完整地描述 K·皮尔逊体系中数字的四个参数分别被称为：

1. 平均数（the mean）——测量值散布状态的中间值；
2. 标准差（the standard deviation）——测量值的散布与平均值偏离有多远；
3. 对称性（symmetry）——测量值在平均值一侧规程的程度；
4. 峰度（kurtosis）——个别的观测值偏离平均值有多远。

[1] 有时叫高斯分布，以纪念曾一度被认为第一个提出它的高斯，不过另外的说法是：并非卡尔·费里德里希·高斯（Carl Friedrich Gauss），而是更早的数学家亚伯拉罕·棣·莫弗（Abraham de Moivre）第一个写下了这一分布的公式。也有充分的理由相信，是丹尼尔·贝努里（Daniel Bernoulli）在那之前就发现了这个公式。这就是当代科学史专家斯蒂芬·施蒂格勒（Stephen Stigler）所说的误称定律（the Law of Misonomy）的例子，数学中根本没有以其发明者命名的东西。

用 K·皮尔逊偏斜分布体系去考虑问题，思路会有一种微妙的转移。在 K·皮尔逊之前，科学所处理的事情都是真实的。开普勒试图发现行星如何在空间运行的数学规律；威廉·哈维的实验打算确定血液如何在某一特定动物的静脉和动脉中游动;化学则处理元素和由元素组成的化合物。然而，开普勒所试图追踪的“行星”实际上是一组数据，用来给地球上的观测者所看到的天空中微弱的光点定位。单匹马身上血液通过静脉流动的实际情形，也许与在另一匹马或者一个人身上所可能看到的不同。没有人能够生产出纯铁的样本，尽管谁都知道铁是一种元素。

K·皮尔逊提出，这些观测到的现象只是一种随机的映像，不是真实的，所谓的真实是概率分布。科学中真实的东西并不是我们所能观测到或能把握到的，它们只是通过用来描述我们所观测事物随机性的数学函数来反应。科学调查中我们真正想确定的，是分布的四个参数。从某种意义上说，我们永远不能确定这四个参数的真实数值，而只可能从资料中估计它们。

K·皮尔逊并没有意识到这关键的一点，他以为，如果我们能够搜集到足够的数据去估计参数，就会得到参数的真实数值。而他的年轻对手费歇尔指出，K·皮尔逊的许多估计方法并不是最优的，在 20 世纪 30 年代末期，当 K·皮尔逊临近他漫长生命的终点之际，一位杰出的波兰年轻数学家耶日·奈曼（Jerzy Neyman）表明，K·皮尔逊的偏斜分布体系并没有包含所有可能存在的分布，许多重要问题不能用 K·皮尔逊的体系解决。

还是让我们离开 1934 年那个被离弃的老皮尔逊吧。回到他三四十岁、精力充沛的时期，那时的他对自己所发现的偏斜分布充满了热情。1897 年，他接管了高尔顿在伦敦的生物统计实验室，带领一支年轻的娘子军（被称为“计算员”），计算高尔顿所积累的人种测量数据的分布参数。在 20 世纪之交，高尔顿、K·皮尔逊和 R·韦尔登（Rerhael Weldon）共同努力，创办了一个新的科学期刊，这将使 K·皮尔逊的观点应用到生物数据上。高尔顿用他的个人财富建立了一个信托基金支持这个期刊。在第一期，编辑们提出了一个雄心勃勃的计划。

生物统计计划

当时，英国科学家中有一位杰出的人物，他就是达尔文，同期的科学家们致力于探索达尔文富有洞察力的见解，高尔顿、K·皮尔逊和韦尔登便是其中相当热心的骨干。达尔文的进化理论认为，生命形式随着环境压力而变化，他提出，变化的环境会给更适应新环境的随机变化提供些许的优势，渐渐地，伴随着环境改变和生命形式继续发生随机转变，新物种将会出现并且更适于在新的环境中生存和繁殖。这一思想被简称为“适者生存”（survival of the fittest）。当恣意妄行的政治学家将其用于社会生活，宣称那些在经济竞争中取得胜利的富人比身陷贫困的穷人更为适于生存时，这一理论对社会就有不好的影响——适者生存理论成了猖狂的资本主义的辩护者，在那里，富人被授予了道义上的特权去鄙视穷人。

在生物科学中，达尔文的思想似乎很有道理。达尔文可以指出相关物种的相似性，作为现代物种从先前物种演化而来的佐证。达尔文表明，物种上些许不同的小型鸟类，即使是生活在孤岛上，也有许多解剖学上的共性。他指出，不同物种胚胎之间的相似性，这包括人类的胚胎，在开始是有尾巴的。

有一件事是达尔文做不到的，那就是他不能给出人类历史的时间框架中，新物种实际出现的例子。达尔文设定新物种由于适者生存而出现，但没有证据，他不得不做的只是展示现代物种很好地适应了它们所处的环境。达尔文的说法似乎只是表明了已知的事情，而且理论本身有一个很吸引人的逻辑结构，但是如果套用犹太人的一句老话就是“举例并不是证明”（For instance is no proof）。

K·皮尔逊、高尔顿和韦尔登打算在他们的新期刊中将这事搞清楚。在 K·皮尔逊看来，

只有概率分布是真实的，达尔文的雀鸟（他在书中用到的一个重要例子）并不是科学调查的对象，而某一种雀鸟的总体随机分布才是这个对象。对某一给定雀鸟种类而言，如果能够测量其全体的喙长，这些喙长的分布函数将有四个参数，这四个参数将是这一种雀鸟的喙长。

K·皮尔逊说，假如存在着某种环境力量，通过提供优越的生存能力，使得某一物种产生某种特定的随机变化，我们也许不能生存得那么久，以看到新物种的出现，但我们能够看到分布的上个参数的变化。在他们期刊的创刊号上，三位编辑宣布：他们的新期刊将从全世界搜集数据，以确定这些分布的参数。最终期望表明，样本参数的变化与环境变化相关。

他们将新期刊定名为《生物统计》(Biometrika)，高尔顿创建的生物统计基金会给予它慷慨资助。由于资金是这样地充裕，以至于该期刊成为世界上第一本印有全彩照片的期刊，甚至还带着画有复杂图画的下班纸折页。期刊以高品质的优质纸印刷，连最复杂的数学公式也展示了出来，尽管那意味着极端复杂和昂贵的排版工艺。

接下来的 25 年里，《生物统计》发表了通讯员们从各地发来的数据：有的深入非洲的丛林，测量原住民的胫骨和腓骨；有的从中美洲的雨林抓到奇特的热带鸟类，测量其喙长；还有的甚至偷盗古墓，揭开死人头盖骨灌铅，以测量其脑的容量。在 1910 年，该期刊发表了几幅全彩照片，画面是俾格米男人裸躺在地上，的生殖器旁还摆着量尺。

在 1921 年，一个年轻的女通讯员朱莉亚·贝尔（Julia Bell）描述了她在试图对阿尔巴尼亚新兵进行人类形体测量时所遇到的困难。她离开维也纳去阿尔巴尼亚一个边远的基地，本以为可以得到讲德语军官的帮忙，当她抵达时才发出，那里只有一个士官能说三句德语。她无所畏惧地拿出了测量所用的铜标尺，通过形体动作让那些年轻人理解她要干什么，直到他们按要求抬起手臂和脚。

对每一组这样的数据，K·皮尔逊和他的计算员们都计算出分布的四个参数，论文将展示最佳分布的图示，并评论该分布与其它相关数据的分布有何不同。回顾过去，很难看出所有这些行动怎样帮助证明了达尔文的理论。浏览《生物统计》的这些作品，我得到这样一种印象：这些工作不久就变成为自身原因而进行努力，除了给特定数据组估计参数外，没有实际目的。

在期刊中还夹杂着其它类型的论文，其中一些涉及理论数学，以处理发展概率分布时遇到的问题。比如在 1908 年，一个不知姓名的作者，以“学生”（“student”）为笔名发表了论文，提出了后来几乎在所有现代科学工作中都有作用的研究成果——“学生”的“t 检验”。接下来的几章我们还会遇到这位匿名的作者，并将讨论他在 K·皮尔逊与费歇尔之间作调解时的不幸角色。

高尔顿死于 1911 年，而韦尔登则于这之前死于阿尔卑斯山的一次滑雪事故。只剩下了 K·皮尔逊这唯一的编辑和信托基金的支配者。在接下来的 20 年中，期刊成了 K·皮尔逊个人的了，期刊发表什么完全以 K·皮尔逊的判断为准，由他确定重要与否。K·皮尔逊为期刊写了很多社论，他让自己丰富的想象驰骋在各个领域。比如，在对一个古老的爱尔兰教堂翻修时，墙壁中发现了一副骨骼，K·皮尔逊通过对这些骨骼的测量和所涉及的数学推理，来确定它们事实上是不是某个中世纪圣徒的遗骨。再比如，一个据称是奥利弗·克伦威尔（Oliver Cromwell）的头骨被发现了，K·皮尔逊以一篇精彩的文章对其进行了研究。该文描述了所知的克伦威尔尸体的下落，并且还将对克伦威尔画像所做的测量结果和该头骨[2]所做的测量进行了比较。在另外一些论文中，K·皮尔逊检验了古罗马各君主的统治期和贵族

[2] 克伦威尔专制政权之后，王室复位。当时英格兰内战的双方达成停战协议，新统治者不得追究克伦威尔的追随者。然而，这项协议只论及幸存者而非死者，于是，克伦威尔和两个判处查理一世死刑的法官的尸体被挖了出来，以弑君罪交送审判。他们被宣判有罪，脑袋被砍下来挂在西斯敏斯特大教堂（Westminster Abbey）的旗杆上，三颗人头挂了几年之后失踪。后来，被认为是克伦威尔的那颗人头出现在伦敦的一家博物馆里，这正是 K·皮尔逊检验的那颗，他的结论是：那的确是奥利弗·克伦威尔的头。

阶级的没落，还涉猎了社会学、政治学和植物学。所有这些，都带有复杂的数学解释。

就在去世之前，K·皮尔逊还发表了一篇题为“论犹太人与非犹太人关系”（On Jewish - Genlile Relationships）的短文。文中他分析了从世界各地收集到的犹太人与非犹太人的人体测量数据，最后得出的结论是：德国国家社会主义（the National Socialists）（正式的名称是纳粹（Nazis））的种族理论纯粹是胡说八道，根本就没有犹太种族（Jewish race）或亚利安种族（Aryan race）那回事。这最后一篇论文与他以前的工作一样，组织清晰，有逻辑性，推理谨慎。

K·皮尔逊运用数学研究了人类思想的许多领域，而很少有人将这些领域视为科学的正宗地盘。浏览生物统计上他所写的社论，你仿佛看到了一个兴趣十分广泛的人，他具有直切问题核心的惊人能力，并能用数学模型去加以处理。还有浏览这些社论，你就像遇上一个意志坚定、主见鲜明的人。说实话，如果不需要与他争辩的话，我想我是很乐意与 K·皮尔逊共处一天的。

K·皮尔逊他们是否证明了达尔文适者生存的进化论理论呢？也许是吧。通过将古墓中头骨的容量分布与现代男女的比较，他们设法证明：经历了几千年深化的人类种群保持了相当的稳定。他们表明：对澳洲原住民的人类学测量与对欧洲人的测量结果有着相同的分布，据此，他们推翻了某些澳洲人关于原住民不是人类的断言。K·皮尔逊从这些工作中发展了一种被称为“拟合优度检验“（goodness of fit test）的基本统计工具，这是现代科学所不可缺少的。它使科学家能够确定一组给定的观测值是否适合于某一特定的数学分布函数。在第 10 章我们会看到，K·皮尔逊的儿子 E·皮尔逊（Eqon Pearson），是如何用这种拟合度检验是否定他父亲所完成的许多项工作的。

随着 20 世纪的来临，《生物统计》中讨论数理统计理论问题的文章越来越多，少量的文章仍停留在处理特定数据的分布。当 K·皮尔逊的儿子 E·皮尔逊接班成为编辑时，期刊的性质就完全转型为理论数学了。时至今天，《生物统计》仍是这个领域中卓越的刊物。

但他们到底有没有证明适者生存这个说法呢？20 世纪初曾经有一个最接近的研究。韦尔登构想了一项宏大的实验：18 世纪英格兰南部瓷器工厂的发展，导致了一些河道被粘土淤塞，普利茅斯（Plymouth）港和达特茅斯（Dartmouth）港也都受到了影响，近陆地区比近海地区淤得更为严重。韦尔登从这些港口抓了几百只螃蟹，分别放入广口瓶中，其中一半用内港的淤泥水，另一半用外港的较干净的水。一段时间后仍有螃蟹存活，韦尔登测量它们的壳，以确定两组螃蟹的分布参数。

正像达尔文所预言的那样，淤泥水中戚的螃蟹在分布参数上有了变化！这是不是证明了进化论呢？不幸的是，韦尔登在写出实验结果前就死了，K·皮尔逊对数据进行了粗略的分析，他描述了这个实验及其结果，但最后的分析却始终没有搞出来。为这项实验提供资助的英国政府要求提供最终报告，但报告了无踪影，韦尔登死了，实验也夭折了。

就生命周期很短的生物，如细菌和果蝇而言，达尔文的理论最终被证明是真实的。用这些物种，科学家可以在较短的一个时间段里完成几千代的实验。现代的 DNA 研究，作为遗传的基石，已经为物种之间的关系提供了更为有力的证据。如果我们假定突变率在过去千万年或更长的时间里保持不变，那么 DNA 的研究可以用来估计灵长类和其它哺乳动物出现的时间框架，至少它经了几百万年。大多数科学家现在都把达尔文的进化论作为正确的东西接受下来。没有其它理论与所知数据吻合的如此之好，于是科学界满足了，原来人们认为需要通过确定分布参数转变来表明较短时间里的进化过程，一日三餐这种观念已经被放弃。

K·皮尔逊的革命所留下来的是这样一个观念：科学的对象并不是不可观测事物本身，而是数学分布函数，以描述与所观测事物相联系的概率。今天，医学研究运用精巧的分布数学模型来确定治疗方法对长期存活的可能效果；社会学家和经济学家用数学分布来描述人类社会的行为；物理学家用数学分布来描述次原子粒子。科学里没有哪一个方面从这场革命中

逃脱。有的科学家宣称，概率分布的使用只是一时的权宜之中，最终我们会找到一种途径回到 19 世纪科学的决定论。爱因斯坦有句名言，他不相信上帝在和宇宙玩骰子，就是这种观点的例子。其他人则相信，大自然基本上是随机的，真实性只存在于分布函数之中。不管一个人的基本哲学是什么，事实仍然是，K·皮尔逊关于分布函数和参数的思想统治了 20 世纪的科学，并在 21 世纪初仍保持着优势。

第 3 章 可爱的戈塞特先生

爱尔兰都柏林的吉尼斯酿造公司（Guinness Brewing Company）是一个声誉卓著的老牌酿造公司，该公司于 20 世纪初开始投资于科学。年轻的吉尼斯刚刚继承这家企业，他就决定雇用牛津和合格大学在化学上顶尖的毕业生，以便将现代科学技术引进到公司的业务中来。在 1899 年，他招募威廉・西利・戈塞特（William Sealy Gosset）进入公司，那是个 23 岁的牛津大学新秀，拥有化学和数学两个学位。戈塞特的数学背景在当时是传统的，包括微积分、天文学和机械式宇宙观下的其它科学分支，K・皮尔逊的创新和后来成为量子力学的萌芽观念，还没有进入大学的课程。戈塞特是由于他的化学专长而被吉尼斯雇用的。对一个酿酒企业来说，要一个数学家又有什么用呢？

戈塞特成为吉尼斯一项很好的投资，他表明自己是一个很能干的管理者，最后他在公司里升任负责大伦敦区业务的主管。事实上，他对本行工艺做出了第一项主要贡献是以数学家的身份来完成的。几年前，丹麦电话公司（the Danish telephone company）是第一个雇用数学家的实业公司，但他们有一个明确的数学问题：制造多大的电话交换板？可制造啤酒又有什么数学问题需要解决呢？

戈塞特在 1904 年发表了第一篇文章，处理的是这样一个问题：麦芽浆准备发酵的时候，需要仔细地测量所用酵母的量，酵母是活的有机体，酵母培育需要保持鲜活，加入麦芽浆前它在瓶中的液体里系列。工人们得到测量清楚某个给定的瓶中有多少酵母，以便决定用多少液体，它们提取一定量的液体，在显微镜下检验，计量他们所看到的酵母细胞数。这种测量有多精确？了解这一点是很重要的，因为麦芽浆中所用的酵母数应该精确地控制。酵母太少，发酵不充分；太多了，啤酒又会发苦。

注意这个问题与 K・皮尔逊对科学的观念是多么的吻合。测量的是样本中酵母细胞的量，但所寻求的真实“东西”是整个瓶中酵母细胞的浓度。由于酵母是活的，而细胞不断地分裂和繁殖，那个“东西”实际上并不存在，在某种意义上，真正存在的是单位液体中酵母细胞的概率分布。戈塞特检验了数据，确定酵母细胞的数量可以用所知的泊松分布（Poisson distribution）[3]来描述，这并不是 K・皮尔逊偏斜分布家族中的一种概率分布。事实上，它是一种只有 1 个（而不是 4 个）参数的特殊分布。

确定了样本中的活酵母细胞数服从泊松分布，戈塞特就能够设计规则和测量方法，从而得到对酵母细胞浓度更为精确的测量。用戈塞特的方法，吉尼斯能够生产质量更稳定的啤酒。

“学生”的诞生

戈塞特想找一份适合的期刊发表这个结果，泊松分布（或相应的公式）已经被发现 100 多年了，过去一直试图在现实生活中寻找实例，其中之一，便是计量普鲁士军队中被马踏死的士兵人数。在酵母细胞计量中，戈塞特有一个清楚的实例，还有对统计分布新观念的重要应用。然而，这违背了公司不准许雇员发表文章的政策。几年前，吉尼斯一位优秀的酿造师写了一篇文章，其中泄露了他们某个酿造过程的秘密成份。为了避免进一步损失，吉尼斯禁止它的雇员发表文章。

戈塞特成了当时《生物统计》编辑之一的 K・皮尔逊的好朋友，而 K・皮尔逊对戈塞特的数学能力印象很深。1906 年，戈塞特说服了他的老板，数学的新思想对啤酒公司是很有用的，并到高尔顿生物统计室在 K・皮尔逊门下脱产学习一年。这之前两年，当戈塞特描述他处理酵母的结果时，K・皮尔逊急于将之付印于他的期刊。他们决定用匿名的方式发表文

---

[3] 就像 S・施蒂格勒所说的误称定律，泊松分布是以 18－19 世纪的数学家 S・D・泊松命名的，但是这个分布却在更早些时候由贝努里家族的一个人描述过。

章，于是，戈塞特的首次发现是仅是以“学生”的名义发表的。

在其后 30 年中，“学生”写了一系列极为重要的论文，几乎所有的都发表在《生物统计》上。从某些方面看，吉尼斯家族已经发现了他们“亲爱的戈塞特先生”违反了公司的规定，一直私下里撰写并发表科学论文。“学生”的数学活动大多是在家里进行，并且是在正常的工作时间之外。戈塞特在公司升迁到了负更多责任的位置，这表明他的副业并没有使吉尼斯公司受损。有这样一种不足为凭的说法：吉尼斯家族第一次知道这件事是在 1937 年，戈塞特突然死于心脏病，他数学界的朋友与吉尼斯公司探讨，想帮助支付其论文集的印刷成本。不管这事真实与否，美国统计学家哈罗德·霍特林（Harold Hotelling）的回忆录里清楚地记载，霍特林在 20 世纪 30 年代后期要与“学生”会谈，安排是秘密的，带有间谍小说的各种情节。这表明“学生”身份的真正确认，对吉尼斯公司仍是个秘密。“学生”在《生物统计》发表的论文涉及理论和实践的尖端问题，戈塞特将非常实际的问题带入有难度的公式，又把结论带回现实实践，后来者便照此办理。

尽管有很高的成就，戈塞特仍是个谦逊的人。在他的信中，人们经常可以发现这样的字眼：“我的研究只是提供了粗浅的想法”；或者，当他的某些发现被给予过多的荣誉，他会说：“费歇尔实际上已经能完成了整个数学结构。”在人们的记忆中，戈塞特是一个和善的、体贴的同事，很在意别人的情感。他去世的时候 61 岁，离开了他的妻子马乔里（Majory）（一个精力充沛的运动员，曾经担任英国女子曲棍球队的队长）、一个儿子、两个女儿和一个孙子，当时他的父母还健在。

“学生”的 t 检验

如果不算别的，所有的科学家都受惠于戈塞特的一篇短文，该文的题目是“平均数的可能误差”（The Probable Error of the Meam），1908 年发表在《生物统计》上。是费歇尔点出这篇杰出论文的一般性意义。对戈塞特来说，有一个特定的问题需要解决，一到晚上，他就习惯性地带着耐心和小心投入于这个问题。发现了结论，他就用其它资料来检查，重新验证他的结果，努力去确认是否遗漏了什么细微的差别，考虑他必须设定哪些假设，并一再重复计算自己的发现。他提前采用了现代计算机基础上才出现的蒙特卡罗技术（Monte Carlo techniques），这是一种一再模拟的数学模型，以确定相关数据的概率分布。然而，当时他没有计算机，只能不辞辛苦地加总数据，从上百个样本中计算平均数，并绘制所得出频率的图表，所有这些都靠手工完成。

戈塞特所专注的特定问题是小样本（small sample）问题。K·皮尔逊计算了某一分布的 4 个参数，这是在单一样本就积累了上千个测量数据的基础上完成的，因为使用了大样本，他设定所得到的参数估计是正确的。费歇尔要证明他的错误。根据戈塞特的经验，科学家很少能三八线以有如此大的样本，更为典型的实验通常能够看到 10 到 20 个观测数据，他还理解到，这种现象在所有的学科中都很普遍。在一封给 K·皮尔逊的信中，他写道：如果我是你遇到的用小样本工作的唯一一人，那你太特异了，在这个题目上我与斯特拉顿（Stratton）（剑桥大学的一位研究员）相伴，他曾经用 4 个样本来做说明。

K·皮尔逊所有的工作都假定：样本足够大，以至于确定参数可以没有误差。戈塞特设问：如果是小样本会怎么样？我们将如何处理自己的计算中肯定会出现的随机误差？

晚间，戈塞特坐在自己的餐桌旁，取出一小组数据，算出平均值和标准差估计值，再将二者相除，并将结果绘在图纸上。他发现这个比率与 K·皮尔逊的四个参数相关，并与 K·皮尔逊的偏斜分布系列中的某一分布相配。他的伟大发现在于：你不必知道原始分布的 4 个参数的确切值。前两个参数估计值的比率有一个可以制表的概率分布，不管数据从哪里来，或者标准差的真实值是多少，计算这两个样本估计值的比率，你就可以得到一个已知的分布。

正如弗雷德里克·莫斯特勒（Frederick Mosteller）和约翰·图基（John Tukey）所指出的那样，没有这一发现，统计分析注定要使用无限次的回归，没有"学生"的 t 检验[4]（这是该发现后来的称谓），分析者将不得不估计观测数据的 4 个参数，再估计这 4 个参数估计值的 4 个参数，接着估计 4 个新估计值的 4 个参数……这样继续下去，没有机会得到最终的结果。戈塞特表明，分析者可以在第一步就停止这种估计。

戈塞特的工作有一个基本的假设，即原始测量值服从正态分布。多年以来，科学家使用着"学生"的 t 检验，许多人渐渐相信，并不需要这项假设。他们经常发现：不管原始测量是否服从正态分布，"学生"的 t 检验都有相同的分布。在 1967 年，斯坦福大学（Stanford University）的布拉德利·埃弗龙（Bradley Efron）证明了这一点，更确切地说，他发现了不需要戈塞特假设的一般条件。

随着"学生"t 检验的发展，我们不知不觉地习惯于统计分布理论的应用，这一理论在科学界广为流传，相伴而来的是更深层次的哲学问题，这就是我们所说的"假设检验"（hypothesis tests）或"显著性检验"（significance tests）的使用。后面我们会剖析这个问题，现在我们只想强调："学生"提供了几乎每个人都使用的科学工具，尽管没有多少人真正理解它。

与此同时，"可爱的戈塞特先生"成了两个长期不和的超级天才——K·皮尔逊和费歇尔之间的中间人。尽管他经常对 K·皮尔逊抱怨他看不懂费歇尔写给他的东西，他还是保持了与两个人的友谊。他与费歇尔的友谊开始于费氏在剑桥大学读本科的时候，那是在 1912 年，费歇尔刚刚成为剑桥大学数学学位甲等及格者（最高的数学荣誉），他的天文学导师[5]介绍两个人认识。当时费歇尔正在研究一个天文学问题，他写了一篇论文，在其中他重新发现"学生"在 1908 年得到的结果。年轻的费歇尔显然不大知晓以前戈塞特所做的工作。

在费歇尔给戈塞特看的这篇论文中，有一个小错误被戈塞特指了出来。当戈塞特回家的时候，他发现费歇尔写的两大页数学论证正等着他。这个年轻人把自己原先的工作又做了一遍，并加以扩充，还批评了戈塞特所犯的一个错误。戈塞特在给 K·皮尔逊的信中写道："附上一封信，它证明了我关于"学生"t 检验的频率分布公式，您是否介意替我看一下。即使我可以理解，超过三维空间我还是觉着不自在。"费歇尔用多维几何证明了戈塞特的成果。

在这封信中，戈塞特说明了自己的如何到剑桥去与朋友会面，而这个朋友恰巧在冈维尔与凯厄斯学院（Gonville and Caius College），是费歇尔的导师，他如何被介绍给这位 22 岁的学生。他接着写道："费歇尔这小子写了一篇论文，提出概率的新标准或诸如此类的东西，看起来不错，但就我所能理解的，是一种不切实际且不大管用的认识事物方式。"

在描述了他在剑桥与费歇尔的讨论后，戈塞特写道：

> 对我们之间的讨论，他的回复是两大页书写纸，上面用最深的墨水写满了他所证明的数学（跟着是一组数学公式）……我看不大懂这些内容，回复他说等我闲下来时准备研究它，实际上我去湖区时随身带着它，可弄丢了。
>
> 现在他将这封信寄给我，我觉得如果它还可以的话，您也许愿意发表这个证明，它是这样的完美和数学化，对某些人也许有吸引力。

K·皮尔逊在《生物统计》上发表了费歇尔的短文，就这样，20 世纪最伟大的天才之一面世了。3 年以后，经过了一连串俯就的信件往来，K·皮尔逊发表了费歇尔的第二篇论文，但事先约定论文须以这种形式出现：它不过是对 K·皮尔逊合作者之一所做工作的细微补充。K·皮尔逊再也没有允许他的期刊发表费歇尔的论文。费歇尔继续在 K·皮尔逊许多最感自豪的成就中挑毛病，而 K·皮尔逊则在稍后几期的《生物统计》中，以社论的方式点出"费

[4] 这是 S·施蒂格勒所说的误称定律的一个例子，戈塞特用字母"z"代表这个比率，许多年后，教材的作者改变了传统上用字母"z"表示正态分布的变量，开始用字母"t"表示"学生"的检验的比率。

[5]这在英国诸如剑桥等学校里是一个惯例，要为每个学生配一个辅导老师，称为学生的导师，其主要职责就是指导学生完成课程的学习.

歇尔先生”或“费歇尔先生的学生”在其它期刊所发表论文中的错误。这些都将是下一章介绍的内容，戈塞特会在以后几章中的某些地方再度出现，作为一个和蔼可亲的良师益友，他帮助年轻男女进入统计分布的新世界。他的许多学生和合作者都对新数学做出了重要贡献。尽管他本人谦逊地表示异议，但戈塞特的确做出了许多影响深远的贡献。

第 4 章 在“垃圾堆”中寻觅

1919 年春天，费歇尔 29 岁，他带着妻子、三个孩子和小姨子，搬到了伦敦北部的一间旧农舍里，那儿靠近罗森斯特农业实验站（the Rothamsted Agricultural Experimental Station）。从许多方面来看，费歇尔的人生在别人眼里是失败的。他在孤单和多病的童年中长大，并有严重的视力损伤。为了保护他的近视眼，医生禁止他在人工灯光下阅读。但他很小就接触了数学和天文学，在 6 岁时他迷上了天文学，七八岁时，他就跑去听由著名天文学家罗伯特·鲍尔（Robert Ball）爵士主讲的通俗讲座。

费歇尔被著名的哈罗公学（Harrow Public School）[6]录取，在那里他的数学是出众的。由于不允许他使用电灯，他的数学导师在晚上教他时，不用铅笔、纸和任何其它视觉辅助品。久而久之，费歇尔发展了一种很强的几何直觉能力。在后来的岁月中，他那非凡的几何洞察力，使他得以解决许多数理统计中的难题。这种洞察力对他而言是那么明显，从而导致他经常不能被别人所理解。在他看来是显而易见的事情，别的数学家往往要花几个月甚至几年的时间去证明。

他于 1909 年进入了剑桥，在 1912 年获得了数学学位甲等及格者的头衔，对剑桥学生来说，这是一个很高的荣誉，要得到它必须通过一系列极为困难的口头和笔头数学考试，一般一年只会有一两个学生成功，有的年份甚至没有人能得到这种头衔。当费歇尔还是本科生时，他就发表了他的第一篇科学论文，其中复杂的迭代公式（iterative formulas）被转换成多维的几何空间形式。在这篇论文中，那些在人们眼里一直特别复杂的数学计算公式被转换成简单的几何形式。毕业后他花了一年时间，研究统计力学（statistical mechanics）和量子理论（quantum theory），到 1913 年，统计革命已经进入了物理学，而新观念已经较为系统地进入这两个领域，并成为正式的大学课程。

费歇尔的第一份工作是在投资公司的统计室，其后他突然离开那里，到加拿大去从事农场工作。后来又在第一次世界大战开始时突然离开农场，回到了英格兰。虽然他被批准入伍，但他那很差的视力使他免于军事服务。战争年代，他在许多公共学校教授过数学，但每一次的经历都比上一次更糟，他对学生们没耐心，因为他们都是不能理解在他看来很明显的事情。

费歇尔与 K·皮尔逊

前一章提到，当费歇尔还是本科生时，就在《生物统计》发表了一篇短文。这使得费歇尔有机会见到 K·皮尔逊，K·皮尔逊将一个困难的问题介绍给费歇尔：确定高尔顿相关系数的统计分布。费歇尔对此作了思考，用几何公式来处理它，不到一个星期就得出了完整的答案。他把结果交给 K·皮尔逊，想在《生物统计》上发表。但 K·皮尔逊不能理解其中的数学，把它转给了戈塞特，而戈塞特在理解上也有困难。K·皮尔逊知道如何就特定的案例得到问题的部分结论，他的方法涉及到大量的计算工作，于是便对生物统计实验室的工人做出安排，让他们去计算出这些明确的答案。在每一个案例中，所得到的答案都更加支持费歇尔的一般性结论。但 K·皮尔逊仍然不发表费歇尔的论文，他要费歇尔做出修改，并降低费歇尔工作的一般性。K·皮尔逊将费歇尔的东西扣了一年多，同时让他的助手（计算员）计算一个庞大的扩展的表，以表明参数值的分布。最后，他发表了费歇尔的成果，但相对于 K·皮尔逊及其助手展示分布表的大块文章来说，费氏的论文只是作为一个脚注。对不经意的读者来说，这样一个结果意味着：K·皮尔逊和他的合作者所做的工作更为重要，那里有大量的数据计算，而费歇尔的数学处理只是一个附属物。

[6] 这是误称定律延伸超出数学领域的一个例子。在英国，像哈罗这样最高级的私立学校，被称为“公学”（Public Schools）。

费歇尔再也没有在《生物统计》上发表过文章，尽管它是这一领域的顶尖级期刊。在接下来的年份里，费歇尔的论文出现在《农业科学期刊》（the Journal of Agricultural Science）、《皇家气象学会季刊》（the Quarterly Journal of the Royal Meteorological）、《爱丁堡皇家学会会刊》（the Proceedings of the Royal Society of Edinburgh）、《心理研究学会会刊》（the Proceedings of the Society of Psychical Research）上，而所有这些期刊与数学研究通常都不怎么搭界。据知情者说，费歇尔作出这样的选择是因为 K·皮尔逊和他的朋友们成功地将费歇尔逐出数学和统计研究的主流。根据其它人的说法，K·皮尔逊吹毛求疵的态度让费歇尔感到自身受到漠视，同时，他也没能够让类似的论文在《皇家统计学会期刊》（the Journal of the Royal Statistical Society，该领域另一份顶尖的期刊）上发表，于是他转而利用其它期刊，有时甚至付钱请他们发表自己的论文。

费歇尔这个“法西斯”!

费歇尔早期论文有一些是高度数学化的。他论述相关系数的文章，也就是 K·皮尔逊最后同意发表的那篇，就充满了数学符号，一个标准页里有一半甚至更多篇幅都是数学公式。但也有一些论文里面压根就没有数学。其中的一篇，他讨论了用达尔文的随机适应理论（Darwin’s theory of random adaptation）来说明最复杂的解剖学结构的方法。在另一篇论文中，他探讨了性别选择进化的问题。费歇尔在 1917 年加入了优生学运动（the eugenics movement），在《优生学评论》（the Eugenics Review）上发表了一篇社论，呼吁转变国民政策“以增加职业界人士和高技能工匠的生育率”，并抵制下层社会的生育率。他在这篇文章中质疑政府为贫民提供福利的政策，认为这会鼓励他们多生育，并将基因传给下一代，而中产阶级对经济安全的关注会导致他们推迟结婚，并节制生育。费歇尔担心，对整个国家来说最终的结果是：为后代选择了“最差的”而不是选择“较好的”基因。优生学问题是通过有选择的系列来改进人类基因库，这成为费歇尔的主要政治观念。在第二次世界大战期间，他被错误地指责为法西斯主义者，并被逐出了与战事有关的工作。

费歇尔的政治见解与 K·皮尔逊不同，后者钟情于社会主义和马克思主义，他同情被压迫者，并喜欢挑战保守的优等阶层。但 K·皮尔逊的政治观念对他的科学研究没有什么影响。费歇尔关注优生学，这导致他将相当大的精力投入到遗传学的数学研究中。当时有一种新观念，认为某种植物或动物的特性可能来自一个单个基因，这以两种形式中的一个就可表现出来。从这种观念出发，费歇尔将格雷戈尔·门德尔[7]的工作大大地推进了，他指出如何估计两个相信基因的彼此影响。

存在着控制生命性质的基因，这一观念是科学中广义统计革命的一个部分。我们观察植物和动物的我，专业上称之为“表型”（phenotypes）。但我们假设这些表形是基因之间交互作用的结果，而这些基因的交互作用又具有不同的概率。我们寻求以这些主要的和不可见的基因方式，来描述“表型”的分布。在 20 世纪后期，生物学家识别出这些基因，以确定它们让细胞制造什么样的蛋白质，我们说起这类事就像真的一样，但我们所观察到的还只是概率的分布，我们所说的基因，即 DNA 链，正是来自于这些分布。

---

[7] 格雷戈尔·门德尔是中欧的一个神父，他的真名字是约翰（Johann，叫门德尔是更大程度的误称），他在 19 世纪 60 年代发表了一系列文章描述豌豆培育的实验。因为它的研究成果与当时的植物学公认的结果不一致，所以没有受到重视。他的研究成果是被剑桥大学一群生物学家重新发现的，威廉·贝特森（Willian Bateson）是他们的领导者，他在剑桥建立了遗传学系。K·皮尔逊所钟情的许多辩论中，其中之一就是表现他对这些遗传学家的研究工作的蔑视，认为他们只是在活的生特体上检查细小的离散变化，而 K·皮尔逊感兴趣的则是重要的分布参数的连续改变，认为它才是进化的真正本质。费歇尔的早期的一篇论文指出，K·皮尔逊的数学公式是可以由贝特森的细小的离散变化推导出来。K·皮尔逊看了后认为这是显然的，并且说费歇尔应该将这篇论文送一份给贝特森，让他知道这个事实。贝特森则说，费歇尔应该将这篇论文送一份给 K·皮尔逊，让他知道这个事实。最后，费歇尔继续贝特森之后 ，担任了剑桥大学遗传学系的主任。

我们这本书说的是总的统计革命，费歇尔在这场革命中起了很重要的作用。他对自己作为遗传学家所取得的成就感到自豪，他的一半以上的成果是与遗传学有关的。现在，我们不再把费歇尔当作一个遗传学家，而主要看他在一般统计技术和观念方面取得的进展。这些观念的萌芽在他的早期作品中就可以发现，但这些观念的全面发展，却是他在工作期间的事，那发生在 20 世纪 20 年代到 30 年代。

《研究工作者的统计方法》

虽然费歇尔在这段时间被数学界忽视了，但他所发表的论文和著作极大地影响了农学和生物学界科学家的工作。在 1925 年，《研究工作者的统计方法》（Statistical Methods for Research Workers）第一版面世。之后，这本书仅英文版就出了 14 个，此外，还有法文、德文、意大利文、西班牙文和俄文的译本。

《研究工作者的统计方法》与这之前的数学著作不同，通常数学著作都有许多定理及其证明，并展开抽象的概念将之一般化，与其它抽象概念联系。如果说这类书中有什么应用的话，也只是放在完整的数学描述和证明之后。《研究工作者的统计方法》从如何利用数据制图及如何读图开始，第 3 页就出现了第一个实例，展示一个婴儿生命头 13 周每一周的重量，这个婴儿就是费歇尔自己的头生子——乔治（George）。接下来的各章描述如何分析数据：费歇尔给出一些公式，列举一些实例，解读这些例子的结果，然后再转到其它公式。书中没有对公式的数学推导和证明，却带有详细的技术说明，并交待如何在机械计算器上应用它们。

尽管，或者说正是因为缺少理论数学，这本书迅速地被科学界采用。它顺应了现实需求，可以把这本书直接交给只受过有限的数学教育的实验室的技工，让他们自己应用。使用这本书的科学家认为费歇尔的主张是正确的，而评论这本书的数学家则对书中未加证明的大胆论述持怀疑态度，许多人弄不明白他是怎么得出这些结论的。

第二次世界大战期间，瑞典的数学家哈拉尔德·克拉美（Harald Cramér）被战争隔绝于国际科学界外，他花了相当多的时间来费歇尔的这本书和所发表的论文，补充了原来缺失的证明步骤，并推导出原来没有的证明。1945 年，克拉美出版了一本书，书名叫作《统计的数学方法》（Mathematical Methods of Statistics），对费歇尔的许多著述给出了正式的证明。不过，克拉美只能对这位多产天才的论述进行选择性的证明，费歇尔的很多著述在克拉美的书中都没有包括进去。克拉美的书被用来教授新一代数学家和统计学家，他把费歇尔著述的“修注”编写成一个标准范式。在 20 世纪 70 年代，耶鲁大学（Yale University）的 L·J·萨维奇（Savage）阅读了费歇尔最初的论文，发现里面有很多东西都被克拉美遗漏了。他还惊讶地看到，费歇尔对后人的工作早有预见，并且已经解决了在 20 世纪 70 年代被认为还没有解决的问题。

但所有这些对 1919 年的费歇尔来说都是未来的事情，当时他正打算放弃不成功的学校老师职业。实际上他刚刚完成一项里程碑意义的工作：将高尔顿的相关系数与门德尔遗传学的基因理论结合在一起。但皇家统计学会和 K·皮尔逊的《生物统计》都拒绝刊登这篇论文。费歇尔听说爱丁堡皇家学会正在寻找适于他们的《交流》（Transaction）上发表的论文，但期望由作者本人支付印刷成本，就这样，费歇尔自费将自己第二项伟大的成果交给这样一个当时并不起眼的期刊发表。

在当时，K·皮尔逊仍对年轻的费歇尔印象很深，他想聘请费歇尔到高尔顿生物统计实验室担任首席统计师，两个人之间的通讯来往是诚恳的，但对费歇尔来说，K·皮尔逊显然是一个主观意志很强并有支配欲的人，所谓首席统计师，充其量不过是在 K·皮尔逊的指令下，从事细节的计算工作。

罗森斯特实验站与农业实验

当时，罗森斯特农业实验站（Rothamsted Agricultural Experimental Station）的所长约翰·罗素（John Russell）爵士也与费歇尔取得了联系。这个实验站是由一个英国的肥料制造商在一个旧农场里建立的。这个旧农场曾属于该肥料公司原来的主人。农场的粘土并不特别适于种植什么作物，但主人发现了如何将石头磨碎与酸混合，生产一种被称作“过磷酸石灰”(Super-Phosphate)的肥料的方法。从过磷酸石灰生产得到的利润用来建立一个实验站，以开发新的人工肥料。90 年下来，这个站进行了许多实验，测试无机盐肥料与不同品第的小麦、黑麦、大麦和马铃薯的不同组合。这积累了一大仓库的数据，有雨量和温度准确的日记录、施肥追肥和土壤测量的周记录、收成的年度记录。所有这些都保存在皮面笔记本中。大多数这样的实验没有产生一致的结果，但这些笔记本被小心地存放在实验站的档案室中。

罗素先生看着积累下来这么多资料，想到也许应该雇个人来看看里边有什么东西，对这些资料进行一次统计整理。他四处询问，有的人推荐了费歇尔。罗素跟费歇尔签了一年的合同，给出了 1000 英磅的酬劳，他只能出这么多了，而且不能保证第二年续聘。

费歇尔接受了罗素的聘任，带着妻子、小姨子和三个孩子来到了伦敦北部的农区。他们租下了实验站旁边的一间农舍，妻子和小姨子打算在那里种种菜园，操持家务，而费歇尔则空上靴子，穿行在农业实验站的田间和 90 年的数据中，做起他后来称之为“在垃圾堆中寻觅”的工作。

第 5 章 收成变动研究

在我担任生物统计学家不久，一次去康涅狄格大学与休•史密斯教授讨论我所遇到的问题，他给了我一份礼物，那是一篇论文的复印件。论文有 53 页长，题目是《作物收成变动研究Ⅲ：降雨量对罗森斯特小麦收成的影响》(Studies in Crop Variation. Ⅲ. The Influence of Rainfall on the Yield of Wheat at Rothamted)。这是一组杰出的数学论文的第三篇，其第一篇 1921 年发表在《农业科学期刊》第 11 郑上。产量变化是实验科学家的大忌，但却是统计方法研究的基本素材。在现代科学文献中，“变动”(variation)这个词已经很少被用到了，它已经被其它术语代替，比方说“方差”(variance)，这个术语与特定的参数分布有关。“变动”对一般的科学用途来说过于含混，但对费歇尔而言，却是合适的，作物产量在年份之间、地块之间的这种变动，正是作者研究的起点，借此，他可以推导出新的分析。

大多数科学论文在结尾都有参考文献目录，一个长长的单子，以确认对所讨论问题曾经有过建树的论文。费歇尔系列论文的第一篇却只有三篇参考文献：其一，指明了 1907 年一次不成功的尝试，打算探讨降雨量与小麦生长的相关性；其二，1909 年以德文写成的，描述了一种计算复杂数学公式最小值的方法；其三，是由 K•皮尔逊发表的一组数表。先前没有什么论文涉足过这一杰出研究系列所涵盖的题目。《作物收成变动研究》是自成一格的，署名的地方写着：罗纳德•A•费歇尔，文学硕士，罗森斯特农业实验站统计实验室，哈盆登(Harpenden)。

1950 年，出版商约翰•威利(John Wiley)征求费歇尔的意见，看他是否愿意从所发表的论文中挑选一些最重要的，好单独形成一本文集。后来这本文集的名称叫做《对数理统计的贡献》(Contributions to Mathematical Statistics)。一打开书，就是费歇尔当时的照片，他一头白发，双唇紧闭，领带稍微有点斜，白胡子梳理得不大好，书中标明费歇尔当时在剑桥大学遗传学系工作。《作物收成变动研究Ⅰ》是该文集中的第一篇文章，作者在文章前面加了一个序言，以明确该文的重要性及其在他全部成果中的地位：

> 早期在罗森斯特的工作中，作者对研究站多年积累下来的大量观察数据，如天气、收成、收成分析等，给予了极大的关注。气象记录在多大程度上能够提供来年收成的预测？对于这类问题，上述数据是有独特价值的。现在这篇文章是用于此目的的系列研究的首篇。

这个系列研究最多有 6 篇论文，《作物收成变动研究Ⅱ》发表在 1923 年，而史密斯先生给我的那篇标号为“Ⅲ”，在 1924 年问世。《作物收成变动研究Ⅳ》则在 1929 年发表。标号为“Ⅴ”的论文没有出现在费歇尔的文集中。在科学史上还很少有这种事件：标题那么不起眼，而其内容却如此重要。在这些论文中，费歇尔开发了用于数据分析的原创性工具，建立了这些工具的数学基础，并描述了如何将它们应用到其它领域中去，包括如何应用到他在罗森斯特所遇到的“垃圾堆”上。这些论文表现了令人眩目的原创性，充满了奇妙的内涵，这足够理论家们在 20 世纪余下来的日子里忙乎的，也许那之后还会继续激发更多的研究。

《作物收成变动研究Ⅰ》

费歇尔系列研究的后两篇文章是有共同作者的，但《作物收成变动研究Ⅰ》却是他独立完成的，那需要大量的计算工作。他的唯一后援是一台名字叫“百万富翁”的计算器，那是一台原始的带有手摇曲柄的机械计算器。如果要算乘法，比方说算 3342 乘 27,先要将转盘放在个位上，设定 3342 这个数字，摇动曲柄 7 次；再将转盘放在十位数上，设定 3342 这个

数，摇动曲柄 2 次，计算方告结束。这架机械叫“百万富翁”，因为它的转盘大得足够容纳以百万计的数字。

为了体会到这篇论文所耗费的气力，我们来考虑一下《作物收成变动研究Ⅰ》中第 123 页的表 7.如果完成一个多位数乘法需要 1 分钟，我估计费歇尔需要大概 185 个小时来完成这张表。这篇论文中有 15 张复杂程度相当的数表，还有 4 张更为复杂的图。只考虑体力劳动本身，准备这些图表至少需要耗去费歇尔 8 个月的时间，而且每天得工作 12 个小时！这还不包括其它工作所花费的时间。比方说：思考理论数学问题、整理数据、设计分析框架、修正不可避免的错误等等。

高尔顿回归思想的一般化

回顾一下高尔顿所发现的“向平均数回归”，他试图找到一个数学公式，将随机事件彼此联系在一起。费歇尔接过高尔顿“回归”（regression）这个词，建立了某个给定地块小麦收成与年份之间的一般数学关系，这个相当复杂分布的参数描述了小麦产量产业化的不同方面。要深入理解费歇尔的数学式，你得有坚实的微积分基础，得对概率分布理论有好的辨别力，还要对多维几何学有感觉，但理解他的结论并不那么难。

他将小麦产量的时间趋势分成几个部分，一个是由于土地退化导致产量稳定地整体性地下降；另一个是长期的缓慢的变化，每个阶段都要花几年时间；第三个是一组更快的移动变化，考虑的是气候在不同年份的差异。自从费歇尔开创性的尝试，时间序列的统计分析在他的思想和方法的基础上，建立了起来，现在我们有了计算机，可以用更巧妙的演算法进行大规模的计算，但基本的思想和方法仍然未变。给定一组随时间波动的数据，我们可以将之分解为不同来源导致的结果。时间序列分析用来检验：美国太平洋海岸拍激的海浪是不是印度洋风暴的起因。这些方法使研究人员能够区分地下核爆破与地震，能够精确地为病理学上的心中节律定位，能够确定环境管制对空气质量的影响，其应用范围还在继续扩大。

农场有一个名称叫“宽田硬”（Broadbalk）的地块，在分析其粮食收成时，费歇尔感到有些困惑，这块地只用了动物粪肥，所以不同年份收成的变动与人工肥料无关。当土壤得自动物粪肥的养分逐渐耗尽，地力退化的长期因素就可以得到解释，同时费歇尔还可以确定不同年份降雨类型不同所带来的影响。那么，什么是缓慢变化的原因呢？从缓慢变化的形态可以看出，在 1876 年产量开始下降，比从另两个因素所能预计的程度还要大，这种下降在 1880 年速度更快了；这种情形在 1894 年开始改善，持续到 1901 年，而后又是下降。

费歇尔发现了带有同样缓慢变化的另一种记录，不过形态是相反的，那是关于麦田里野草的。1876 年后，野草蔓延得越发严重，而到了 1894 年突然开始消失，只是在 1901 年又开始茂盛起来。

后来发现，雇用小男孩到地里去拔草，在 1876 年以前是通告的做法。在英格兰的大地上，下午经常可以看到瘦弱的小男孩穿行于田间，不停地拔草。到了 1876 年，教育法（the Education Act）使得上学带有强制性，田间小男孩的大部队开始不见了。而 1880 年第二部教育法通过，对致使孩子辍学的家长施以罚款，田间剩下的男孩也离开了。没有了拔草的小手，那些野草就又茂盛起来了。

那么，在 1894 年又是什么事情发生，使得趋势逆转了呢？在罗森斯特附近有一所女子寄宿学校，新校长约翰·劳斯（John Lawes）相信，充满活力的户外活动有助于他那些年轻的被托管人的健康。他和实验站的头儿一起安排，让这些年轻姑娘在周六和傍晚出门，到地里去拔草。1901 年劳斯去世后，这些小姑娘恢复久坐的习惯，多是在户内活动，野草也就又回到了“宽田埂”。

随机化控制实验

第二篇研究收成变动的论文也是发表在《农业科学期刊》上，时间是 1923 年。这篇论文并不处理罗森斯特过去实验所积累下来的数据，取而代之的是新实验：一组不同的人工肥料组合对不同品种马铃薯的影响。费歇尔到了罗森斯特后，实验有了明显的改善。不再将某种实验的人工肥料用于整个农场，现在他们把土地划成小的地块，每个地块进一步区分作物的行，地块中的每一行都给予不同的处理。

基本的想法是简单的，之所以简单，那是因为一经费歇尔提出后，它就简单了，但这之前却没有人想到它。任何人观察土地上的作物时，都会很明显地感到有的地块土质好于其它地块。在某些角落，作物长得又高又密，而其它角落，作物则又细又稀。这可能是由于排水方式、土壤类型的改变、未知养分的出现、多年生野草的抵制，或者一些其它未能预见的原因。如果农业科学家要测试两种人工肥料间的区别，他可以将一种施于地块的其它角。但这会将肥料的效应与土壤或者排水等的效应混淆在一起。如果试验在相同的地块不同的年份进行，又会把肥料的效应与气候变化的效应相混淆。

如果同一年里，在相同作物上进行肥料的比较，土壤的差别就会减到最低程度，但他们仍然存在，因为所处理的作物不会有绝对相同的土壤条件。如果我们使用足够多的成对比较，在某种意义上，土壤差异所造成的区别就会被平均掉。假定我们要比较两种肥料，其中一种磷肥的含量是另一种的两倍，我们将地分成小块，每一块有两行作物。我们总是将磷肥多的施于北边这行，南边的那行则施磷肥少的。做到这里，反对的声音就会出来了。如果土壤的肥力梯度（fertility gradient）由北向南，那么北边这行的土质就会比南边那行稍好一点，土壤差异的影响就不会被平均掉。

别急！我们正要做调整，在第一个地块，我们把磷肥多的施在北边，到了第二地块，它将被施在南边，就这样来回调整。我的读者中可能有的已经画出地块的草图，将施磷肥较多的行标上了记号。它会指出，如果肥力梯度从西北向东南，施以额外的磷肥的行将总是比别的行土质好。也会有人指出，如果肥力梯度从东北向西南，结论正好相反。好啦，另一个读者发问了，到底谁对了呢？肥力梯度究竟如何分布？我们的答案只能是：天晓得！肥力梯度这个概念是抽象的，当我们选择从北到南或从东到西时，肥力的真正形态可能以非常复杂的方式上下变动。

我可以想象得出来，当费歇尔提出小地块定型处理将得到更为细心的实验时，罗森斯特的科学家们之间也会有这样的讨论。我也可以想象，当讨论集中到如何确定土地的肥力梯度时，费歇尔笑咪咪地坐在一边，听任他们卷入复杂的争论。他已经考虑过这些问题，并有了简明的答案。了解他的人这样描绘费歇尔：即使是争论触及到他，他仍是静静地坐在那里，吞云吐雾，等等容他给出答案的时机。终于，他拿开嘴上的烟斗，说道："用随机的方法吧！"

费歇尔的变异数分析

的确简单，科学家以随机的方式设计同一地块里不同行家作物的处理，由于随机处理没有固定模式，任何可能的肥力梯度结构都在平均意义上被抵消掉了。费歇尔猛地起身，兴奋地在黑板上写了起来，一行又一行数学符号，手臂在数学公式间挥来挥去，抵消公式两端相同的因子，最后出现的可能是生物科学中最为重要的工具了，在精心设计的科学实验中，如何分解各种不同处理的效应？费歇尔将这个方法称作"方差分析"(ananlysis of variance)。在《作物收成变动研究 II》中，方差分析第一次面世。

《研究工作者的统计方法》列出了方差分析某些例子的计算公式，但在这篇论文中，他给出了公式的数学推导，不过推导过程还没有详尽到学院派数学家满意的程度。所展示的代

数式是为了这样一种特殊情形：比较三种类型的人工肥料、十种不同品种的马铃薯和四个地块。如果比较两种人工肥料、五种马铃薯，或者六种人工肥料、一种马铃薯，则需要几个小时的艰苦工作，以调整出新的代数式。至于搞出适合所有情形的一般公式，就需要更多的数学工作了，恐怕得出几头汗水吧！当然，费歇尔知道一般公式，对他来说，那是如此的明显，以至于没有必要展示它们。

难怪与费歇尔同时代的人对这个年轻人的成果感到困惑！

《作物收成变动研究Ⅳ》介绍了费歇尔年说的"协方差分析"(analysis of covariance)，这是一种因素分解的方法，存在着并非由实验设计而来的条件，它们的效应是可以测量的。当时某医学期刊上发表了一篇论文，描写了针对性别和体重所做调整的治疗效应，用的实际上就是费歇尔在Ⅳ号论文中开创的方法。Ⅳ号论文提出了实验设计的精华，Ⅲ号论文，即史密斯教授推荐给我的那篇，将在本章后边一点儿再讨论。

自由度

1922 年，费歇尔终于第一次在《皇家统计学会期刊》上发表了他的论文。那是一篇短文，适度地指出了 K·皮尔逊公式中的一个错误，许多年后谈到这篇论文，费歇尔写道：

> 这个短文，尽管带着稚气，不那么完整，但却是破冰之举。它是带试验性质的，并且零零碎碎的，有的读者会因此而气恼，可他们不要忘了，它不得不在批判者中找到发表的渠道。对这些批判者来说，摆在第一位的就是绝不相信 K·皮尔逊的成果需要改正，即使是承认了这一点，他们也觉得这事轮不到别的人。

1924 年，费歇尔得以在《皇家统计学会期刊》发表别一篇论文，更长一些，更为一般化。后来在一份经济学期刊上，他对这篇论文及相关的另一篇做了如下的评论："（这两篇论文）要借助于'自由度'（degrees of freedom）这个新概念，来调和由不同作者观测到的有差异和表现异常的结果……"

自由度这个新概念是费歇尔的发明，这直接得益于他的几何洞察力和将数学问题置于多维几何空间的能力。所谓"异常的结果"出现在一本不大引人注目的书里，那是一个名叫 T·L·凯利（T. L. Kelley）的人在纽约出版的。凯利发现有一些数据用 K·皮尔逊的公式似乎不能得出正确的答案。看来只有费歇尔注意到了凯利的这本书，凯利的异常结果只是作为一个跳板，借此费歇尔彻底推翻了 K·皮尔逊另一个最引以为自豪的成就。

《作物收成变动研究III》

《作物收成变动研究》第三篇发表在 1924 年的《伦敦皇家学会哲学学报》（the Philosophical Transactions of the Royal Society of London）上，它是这样开头的：

> 现在就气候对农作物影响而言我们知之甚少，尽管它对一个大的民族产业如此重要。课题的难解，部分地可以归于问题本身固有的复杂性，还有……缺少在实验或者自然产业条件下所取得的数据……

按下来就是长达 53 页的精彩论述，其中包含着现代统计方法的基础，任何学术领域，包括经济学、医学、化学、计算机科学、社会学、天文学、药学，只要是需要建立大量相互关联原因的相关效应，就需要应用这些方法。论文中包含了特别精巧的计算方法（回想一下费歇尔只有那台手动的"百万富翁"用来工作），及如何为统计分析组织数据的良策。我将永远感激史密斯教授，他把这篇文章推荐给我，每次我读起它都会有新的收获。

《费歇尔文集》有五卷本，第 1 卷以 1924 年的论文作为结尾，靠近卷尾的地方，有一张费歇尔 34 岁时的照片，他双手交叉在胸前，胡子修理得挺整齐的，眼镜也没有以前照片

中的那么厚，神情安详而自信。在这之前的 5 年里，它在罗森斯特建立了出众的统计部门，雇用了像弗兰克·耶茨（Frank Yates）那样的合作者。在费歇尔的鼓励下，耶茨将继续对统计分析的理论和实践做出贡献。除了少数例外，K·皮尔逊的学生大多默默无闻，当他们在生物统计实验室工作的时候，只能协助 K·皮尔逊而不能超越他；反观费歇尔，他的多数学生响应了所得到的鼓励，独辟蹊径，赢得了辉煌。

1947 年，英国广播公司（BBC）广播网邀请费歇尔做一个系列讨论，阐述科学的本质与科学研究，在其中一讲的开头，费歇尔这样说道：

> 科学生涯从某些方面看是奇异的，科学存在的理由，是要增加对自然知识的认知。有时候，虽然会有这种认知的增加，但是这个过程不是顺利的，并且是令人感到痛苦的。理由是：人们不可避免地会发现以前所得出的观点，至少在一定程度上，明显是过时的或者错误的。我想大多数人可以认识到这一点，如果已经教授了 10 年左右的东西需要修正，他们会以下面的态度加以接受。但有一些人绝对不能接受，就好像打击了他们的自尊心，甚至是对他们一直把持的私有领地的侵犯。他们必然做得像知更鸟和苍头燕雀寻亲残忍，在春天里我们可以看到，当自己的小巢被冒犯里，它们所表现出的愤怒反应。我并不认为能对此做什么补救。这是科学过程中所固有的特性。但年轻的科学家应该得到提醒和指导，当他们奉献出珍宝去丰富人类的宝库时，必然有人会拒绝他或排挤他。

第 6 章 “百年不遇的洪水”

有什么能比百年不遇的灌水更让人无法预料的呢？洪水奔腾肆虐，泛滥成灾，惨烈至极，确实是百年难得一遇。谁能为这样的突发事件制定防范计划呢？像这样罕见的洪水，我们又怎么能估计其洪峰会高达多少呢？如果说现代科学有统计模型能用来处理观测数据的分布，那么，对这种未曾发生过，或者即便发生，也是百年才发生一次的大洪灾，又该如何用统计模型来分析呢？伦纳德·亨利·凯莱布·蒂皮特（Leonard Henry Caleb Tippett）找到了答案。

L·H·C·蒂皮特 1902 年出生在伦敦，并在伦敦的帝国学院（Imperial College）读物理学，1923 年他从帝国学院毕业。蒂皮特曾说过，他之所以被物理学所吸引，是因为物理学对“精确测量的坚持，……和当时科学辩论的那种学院式方法。回顾自己年轻时的激情，他继续说：“我们通常是把一个假设视为对或错，并把至关重要的实验当作加深认识的主要手段。”当他有机会做实验时，他发现实验的结果与理论预测的结果从未有过精确的一致。依据他自己的亲身体验，他说：“我发现最好是去改进抽样技术（这里他指的是统计分布），而不是丢弃理论。”蒂皮特认识到，他如此钟爱的理论所提供的信息仅仅是有关参数的，而不是具体的观测值。

这样，L·H·C·蒂皮特（当他因发表的文章而著称的时候）通过他自身对实验的理解，开始融入统计变革中来。从帝国学院毕业后，他在英国棉花工业研究协会任统计师。人们通常称这个研究协会为雪莉研究会（Shirley Institute）。该研究会的研究目标主要是利用现代科学方法改进棉线与棉布的生产工艺，其中，他们所遇到的最棘手的问题之一是新纺棉线的强度。因为，即使是在相同条件下纺出来的棉线。其强度也存在很大的差异。蒂皮特非常仔细地做了一些试验，在显微镜下观察那些经过不同拉力抻拉后的棉线，结果他发现，棉线的断裂取决于棉线中最脆弱的纤维的强度。

居然是那些最脆弱的纤维！那么，怎样建立一个描述最脆弱的纤维强度的数学模型呢？由于无法解决这个难题，蒂皮特提出申请，并于 1924 年获准，到伦敦的大学学院高尔顿生物统计实验室（the Galton Biometrical Laboratory），在 K·皮尔逊手下进修一年。关于这段经历，蒂皮特这样写道：

> 在大学学院度过的那段时光让我刻骨铭心。K·皮尔逊是位非常了不起的人物，并且我们也能深切地感受到他有多了不起。他工作勤奋、充满热情，而且关于激励他的下属和学生。我在那里进修的时候，K·皮尔逊依旧在做研究，并且经济热情洋溢、充满激情地出现在课堂上，讲解他刚刚研究出来的最新成果。那些年，虽然他的研究方式有点过时了，但他讲的课仍旧激动人心。……有一门他讲授的课程“17 和 18 世纪的统计学史”，就是他研究兴趣广泛的一个典型代表。……他还是个精力充沛的辩手，……他出版了一套丛书，就叫做《一个好问者与他的问题》（Questions of the Day and of the Fray）……昔日充满活力与辩论的影响随处可见。系里的墙上装饰着格言与漫画，……有一幅关于“油嘴山姆”（Soapy Sam）的讽刺漫画，画的是那位大名鼎鼎的威尔伯福斯大主教（Bishop Wiberforce），漫画作者名为“间谍”。1860 年在英国科学促进协会的会议上，这位大主教曾就达尔文的进化论与 T·H·赫胥黎（T. H. Huxley）进行过一场短兵相接的舌战。此外，还陈列了一些在过去数十年内发表过的出版物，看这些出版物的题目会给人留下一个深刻的印象，那就是该系的研究兴趣十分广泛。如“人类遗传宝典（人的身体、精神与病理性的谱系）”以及“达尔文进化论、医学发展与优生学”。在一次全系 的年度聚餐会上，K·皮尔逊用一种曾为高尔顿提供年度工作报告的方式来总结这一年的工作，就好像高尔顿依然健在，这让我们大家想起他与高尔顿之间非常密切的合作。于是我们共同举杯，“为已故去的生物统计学前辈干杯。”

这是 K・皮尔逊一生中还活跃的最后几年，此后，他的科学成就大部分都被费歇尔和自己的儿子扫进了垃圾桶，成了被遗忘的思想。

尽管在 K・皮尔逊在实验室里有那么多激励，尽管蒂皮特在进修期间学到很多数学知识，然而有关最不牢固的纤维强度的分布问题依然没有解决。回到雪莉研究所之后，蒂皮特发现了学期在最伟大的数学发现背后的一个简单的合乎逻辑的原理，他找到了一个看似简单的方程式，它能把样本数据的分布与极值（extreme values）的分布连在一起。

能写出方程式是一码事，解出这个方程则是另外一码事。为此，他去请教 K・皮尔逊，但没有获得丝毫的帮助。在过去的 75 年里，工程学专业已经积累了大量的方程及其解，这些都能在那些大部头的概览中查到。然而，在这些概览中蒂皮特却找不到他的方程式。

于是，他采用了一个做法，就像一个可怜的高中生做代数题一样，先猜了一个答案，并把答案代进方程式，居然解出了这个方程。但是，对这个方程式而言，这是唯一解吗？对他的问题而言，这恰好是“正确”答案吗？为此，他请教了费歇尔，费歇尔不仅能导出蒂皮特所猜的解，而且还给出了另外两个解，并指出，这些就是仅有的解。这就是所谓的“蒂皮特的三条极值渐近线”（Tippett’s three asymptotes of the extreme）。

极值分布

知道极值分布有什么用处？如果我们知道极值分布与正常值的分布之间的关系，就可以记录每年洪峰的高度，并预测百年不遇的洪灾发生时最有可能的洪峰高度。能够这样做的原因是，每年的灌水测量值给我们提供了足够的信息，用它就可以蒂皮特分布的参数。因此，美军工兵署（USACE）就能计算出在河上究竟该筑起多高的堤防，环保署就能规定气体排放标准来控制工业烟囱废气突然排放的极值，棉纺工业就能确定在棉线生产中究竟有哪些因素会对最脆弱的纤维强度的分布参数产生影响。

1958 年，当时在哥伦比亚大学（Columbia University）任工程学教授的埃米尔・J・冈贝尔（Emil J. Gumbel），出版了那本关于极值的权威教材，书名是《极值统计学》（Statistics of Extremes）。自那时起，由于他的思想已经扩展到许多相关的地方去，极值理论方面的建树就很少了。然而，冈贝尔的这本教材里包含了一个统计学家在处理这类问题时必备的一切知识，书中不仅包括蒂皮特的原创研究成果，而且还包括后来对该理论的精心的改进，其中有很多都是冈贝尔自己的研究成果。

政治谋杀

冈贝尔的一生富有传奇性。在 20 世纪 20 年代末至 30 年代初，他是德国一年大学里资历尚浅的一名教师。从他早期发表的论文中看得出来，他是个极具潜能的人，只是当时还没有机会得到一个令人尊敬的地位罢了。同样，他当时的职位也远算不上稳固，是否有能力养家糊口，还取决于政府那些权威的随心所欲。当时，纳粹在德国境内已经渐趋猖獗，国家社会主义工人党[8]虽然是正式的正常组织，实质上却是由一群歹徒纠集而成的。俗称“褐衫队”（Brown Shirts）的纳粹冲锋队是一个专门从事恐吓与胁迫、恣意暴力和谋杀来执行纳粹党意志的暴徒组织。任何批评纳粹党的人都会遭到暴力攻击，而且通常就发生在城市的大街上，以杀一儆百。冈贝尔有个朋友就是这样在光天化日之下曹到攻击并被公然杀害的。照理说，会有许多目击证人可以指认凶手，但法院往往宣称罪证不足而使纳粹突击队逍遥法外。

冈贝尔曾参加过一场审判，他亲眼目睹了法官全然无视任何证据，恣意裁决，纳粹党徒则在法庭上肆无忌惮地狂呼。对此，冈贝尔惊骇万分。于是，他开始着手调查那些凶手公然

[8]即纳粹党。——译者注

行凶的其他案例，结果没有一例被判有罪。最终他得出结论：司法部门已经被纳粹党人所控制，很多法官要么是纳粹的支持者，要么干脆就是纳粹所雇佣的。

冈贝尔搜集了许多案例，走访证人，证明判决那些凶手无罪是错误的。1922 年，他出版了《四年的政治谋杀》（Four Years of Political Murder）一书，把他搜集调查的结果公之于众。由于发现很多书商根本不敢销售他的书，他不得不亲自去为自己的书安排发行分销。与此同时，他还在继续搜集案例，并于 1928 年又出版了《政治谋杀的原因》（Causes of Political Murder）一书。此外，他还设法成立一个反纳粹的政治团体，但是他的多数学术界同事太害怕了，甚至那些犹太籍的朋友们都吓得不敢参加。

1933 年纳粹党取得了政权，当时冈贝尔正在瑞士参加一个数学会议。他本打算立即赶回德国去与这个新政权做斗争，但朋友们极力劝阻了他，因为只要他一越过边境，就会立刻遭到逮捕，并被处决。在纳粹掌权的最初阶段，在这个新政府还没来得及控制所有的出入境事务之时，少数犹太籍教授，如德国的顶尖的概率论大师里夏德·冯·米泽斯（Richard Vin Mises），他们已经预料到即将发生的灭顶之灾，提前逃离了德国。冈贝尔的朋友也趁这段有利的混乱时机，带着他的家人离开了德国。他们跑到法国暂避一时，但是，1940 年纳粹又入侵了法国。

冈贝尔与家人继续逃往尚未沦陷的法国南部。当时统计法国的是纳粹扶植的傀儡政府，对德国惟命是从。像冈贝尔这样的德国民主党人已经是危在旦夕，因为他们都被列入了叛国者的黑名单，纳粹要求法国政府将这些人移交过去。除了冈贝尔，滞留在法国马赛的德国逃亡者还有德国作家托马斯·曼（Thomas Mann）的哥哥海因里希·曼（Heinrich Mann）、犹太裔小说家利翁·福伊希特万格（Lion Feuchtwanger）。当时驻马塞的美国领事海勒姆·宾厄姆四世（Hiram Bingham IV）违反美国国务院的规定，擅自给这批德国流亡者发了签证。宾厄姆为此受到华盛顿的谴责，最终由于此举而丢掉了他在马赛的职位，但宾厄姆毕竟尽他所能拯救了很多人，这些人如果留在纳粹统计下，将必死无疑。冈贝尔与家人到了美国之后[9]，在哥伦比亚大学谋到一个职位。

数学著述有很多种不同的写法。有此所谓“权威”教科书，内容贫乏、苍白、毫无生气，提出一系列的定理及证明，却几乎引不起读者的任何兴致；有此书通篇是从假设到结论的证明，玄虚而艰涩；而有此权威的教科书，则由始至终充满了精彩的证明，其中的数学推导过程被浓缩成看上去很简单的步骤，按照这些步骤可以毫不费力地得出最终结论；还有极少量的权威性的教科书，作者试图在书中把问题的背景和思想都交代清楚，不仅记述了学科的历史渊源，而且所举的例子也取自生动的现实生活。

最后一类所说的权威性教书的这些牲恰是对冈贝尔的《极值统计学》一书的真实描述。这本书提供了大量有关该学科发展的参考，是对一个高难学科的最为明晰的解释。该书的第 1 章“目录与手段”介绍了该书的主题以及在其他章节中必须理解的数学的发展。这一章本身就是对统计分布理论的数学知识的最卓越的介绍。它的设计思想是让那些只读过大学一年级微积分的学生能看得懂。我第一次读这本书的时候，尽管已经拿到了数理统计博士学位，还是从第一章中获准颇多。作者在前言中谦虚地说：“我期望，而决不是预料，本书的写作能使人类从中获益，哪怕是因为对科学进步的微不足道的贡献。”

这本书的贡献决不能称之为“微不足道”，它是由 20 世纪一位大师级的教师矗立的一座丰碑。集非凡的胆识与杰出的表达能力于一身，把最难理解的思想以条理清晰、简洁精炼的方式表达出来，埃米尔·J·冈贝尔正是这些极为罕见的杰出人才当中的一位。

[9] 1966 年冈贝尔去世的时候，他的文章都交给了纽约的利奥·贝克研究所（Leo Baeck Institute），该研究所最后公开了冈贝尔反纳粹活动的 8 卷微缩胶片资料，资料名称为“埃米尔·J·冈贝尔文集：一个反纳粹学者在魏码政权和流亡期间的政治论文”（The Emil J. Gumbel Collection, Political Papers of an Anti-Nazi Scholar in Weimar and Exile）。

第 7 章 费歇尔获胜

英国皇家统计学会（The Royal Statistical Society）拥有三种可以发表论文的学术期刊，每年学会还主办学术会议，会上邀请演讲者介绍他们最新的研究工作。论文要在这些期刊上发表是相当困难的，必须经过至少两位评阅人的审查，看内容是否正确，而且编辑与主编都必须认为该篇论文代表了当时在自然科学领域的显著进展。但是，与应邀在大会上演讲相比，在学会期刊上发表论文就显得容易多了。大会演讲，这只是留给那些在统计学领域里最杰出的研究人员的一种荣誉。

每一次应邀演讲结束之后，按照学会的惯例，都会组织一场与会者参加的讨论会。由于特邀的会议来宾已经预先拿到了将在大会上演讲的论文副本，因此他们的讨论常常不但详尽，而且一针见血。之后，这篇论文连同讨论会上对论文的评论意见都会发表在《皇家统计学会期刊》上。

这种讨论会，正如在期刊上所展现的，有一种非常程序化的英国风格。大会主席（或某个被指定的人）首先站起来提议向演讲人表示感谢，紧接着陈述他的评论。随后，一位事先指定的皇家统计学会的资深会员直立再次提议表示感谢，并随之发表他的评论。接下来，学会中一些最负声望的会员一个接一个地相继站起来发表他们的评论。除了学会的会员之外，大会还经常邀请一些来自美国、英联邦和其他国家的来宾，也请他们发表评论。演讲人再对所有的评论做出回应。最终，学会期刊允许评论人及主讲者对属于他们自己的那部分文字进行编辑之后才正式发表。

1934 年 12 月 18 日，在学会会议上宣读这样一篇论文的无上荣誉赋予了理学博士、英国皇家学会会员费歇尔教授。经过了 20 世纪 20 年代事实上的孤立之后，费歇尔的天赋终于得到了公认。我们在前几章里读到他的时候，费歇尔的最高学位还只是个理学硕士（M.S.），他的“大学”也不过是伦敦郊外一个偏僻的农业试验站。到 1934 年，他又获得了一个理学博士学位，并且当选为威望很高的英国皇家学会的会员（缩写为 F.R.S.）。直至此时，皇家统计学会才终于承认了他作为这个领域中的领军人物，应该占有一席之地。因为这项荣誉，费歇尔在大会上宣读了一篇论文，题为《归纳推理的逻辑》（The Logic of Inductive Inference）。大会主席是皇家统计学会当时的会长。皇家学会会员 M·格林伍德（M. Greenwood）教授。费歇尔的论文印出来共计 16 页，另外还呈上一份结构严谨、条理清晰的论文摘要，概括了他最新的研究工作。第一位发言的评论人是 A·L·鲍利（A. L. Bowley）教授，他站起身来提议表达谢意，接着发表了他的感言：

> 我很高兴有这样一个机会向费歇尔教授表示感谢。不仅是因为他刚才为我们宣读的论文，更重要的是因为他对统计学的全面贡献。今天借此良机，我谨代表所有我熟悉的统计学家，对他带给统计学研究的无与伦比的热忱，对他提出的数学工具的威力，对他在这里、在美洲和在世界各地的广泛的影响力，以及对他深信做为数学的正确应用所发挥的激励作用表示钦佩之意。

K·皮尔逊当时不在讨论者之列。此前 3 年，他已从他任职的伦敦大学退休。在他的领导下，高尔顿生物统计实验室已经成长为大学里一个正式的生物统计学系。他退休后，该系一分为二，费歇尔受命担任其中之一的优生学系的系主任，另一个则是规模缩小了的生物统计学系，系主任由 K·皮尔逊的儿子 E·皮尔逊担任，同时他还负责高尔顿实验室的工作，并兼任《生物统计》杂志的编辑。

费歇尔与小皮尔逊的私交不大好，这完全是费歇尔的过错。他对 E·皮尔逊的态度带着显而易见的敌意。小皮尔逊这位温文尔雅的先生，一则是代父受过，因为费歇尔不喜欢他的父亲老皮尔逊；二则是代合作伙伴耶日·奈曼受过，费歇尔特别讨厌奈曼（奈曼与 E·皮尔逊的合作将在第 10 章介绍）。尽管如此，小皮尔逊倒是极其尊重并高度评价费歇尔的工作。

多年后他曾写道，他早就习惯了费歇尔从不在著述中提到他的名字。但是，尽管两人之间关系紧张，尽管两系之间存在着争夺权限的纠纷，费歇尔和 E·皮尔逊都清寒是派学生去听对方的课，竭力避免公开的冲突。

至于 K·皮尔逊，此时的他已被学生们称之为“老家伙”了。他拥有一个研究生助手，并保留着一间办公室，但他的办公室无论离两个系的办公地点还是离生物统计实验室，都有一段距离。从美国来的邱吉尔·艾森哈特跟随费歇尔和 E·皮尔逊进修一年，这期间他曾想去拜访 K·皮尔逊，但他的同学和系里的同事都极力劝阻他。问他，为什么不去请教才华横溢的费歇尔，竟然想去看 K·皮尔逊？去看那个老家伙能有什么新的收获？令艾森哈特万分遗憾的是，他在英国期间未曾去拜访 K·皮尔逊，而就在那一年老皮尔逊去世了。

费歇尔学派与皮尔逊学派：两种统计观

哲学上的分歧使费歇尔与 K·皮尔逊在研究统计分布的方法上分道扬镳。K·皮尔逊把统计分布视为对他所分析数据的集合的真实描述。而按照费歇尔的观点，真实分布只是一个抽象的数学公式，搜集的数据只能用来估计这个真实分布的参数。既然所有的估计都有误差，那么费歇尔提出来的一些分析的手段，可以把这种误差的程度降到最低，或者可以更经常地得出比其他任何手段都更接近真实分布的答案。

在 20 世纪 30 年代，看上去是费歇尔在这场辩论中获胜了，但到了 70 年代，皮尔逊学派的观点东山再起。直到写这本书时，统计学界在这个问题上已经分裂成两派，尽管 K·皮尔逊本人几乎不接受他的天才继承者们的观点。费歇尔用他条理清晰的数学头脑廓清了残存在 K·皮尔逊观点中大量的混淆，正是这些混淆使得 K·皮尔逊没有意识到自己观点的深层本质，因此，后来东山再起的皮尔逊方法已经无法回避费歇尔的理论成果。当把统计模型应用于现实时，存在着一些很严重的问题。因此，本书打算在多处探讨这些哲学问题，这里就是其中的一处。

K·皮尔逊把测量值的分布视为一个真实的存在。在他的方法里，对于一个给定的情况，有一个庞大的然而却是有限的（finite）测量值的集合。在理想情况下，科学家会搜集所有的这些测量值，并确定其分布参数。如果无法搜集到全部测量值，那么就搜集一个很大的并且具有代表性的数据子集（subset）。由这些大量的、且具代表性的子集计算出来的参数会与完备集合的参数相同；此外，那些用来计算完备集合参数值的数学方法也适用于有代表性的子集的参数估计，而不会有严重的误差。

但依照费歇尔的观点，测量值是从所有可能出现的测量值中随机选取的，依据随机选取的数据计算得出的一个参数的任何估计值，其结果本身也具有随机性，因此，也会服从一种概率分布。为了能清楚地区分参数的估计值与参数本身这两个不同的概念，费歇尔把这个估计值称为“统计量”（statistic）；不过现代术语往往称其为“估计量”（estimator）。假设我们有两种不同的方法可以得到一个统计量，以估计某个特定的参数。例如老师想了解一个学生对知识掌握到什么程度（参数），就在全班进行了几次测验（测量），并且计算出测验的平均分数（统计量）。那么，究竟是用中位数（median）作统计量“更好”呢，或是取这几次测验中的最高分与最低分的平均值“更好”呢，还是去年最高分与最低分然后把其余的测验成绩加以平均“更好”？

既然统计量是随机的，那么讨论这个统计量的某个值的准确性到底有多大是毫无意义的。我们需要的是一个判别的准则，这个准则以统计量的概率分布为依据，就像 K·皮尔逊所指出的那样，对一组测量进行估计，必须根据它们的概率分布，而不是根据个别观测值。评判哪一个是好的统计量，费歇尔提出了如下三个准则：

一致性（consistency）：得到的数据越多，计算出来的统计量接近参数真值的概率就越

大；

无偏性(unbiasedness):如果用很多组不同数据集多次测量某一特定的统计量，那么该统计量的这些测量值的平均数应该近似于这个参数的真值；

有效性（efficiency）：统计量的值不会完全等于该参数的真值，但是用来估计一个参数的大多数统计量应该与真值相去不远。这些阐述似乎有点含混不清，这是因为我在竭尽全力地把一些本来精确的数学公式，用一些一般性的文字表述出来。实际上，费歇尔的这些准则都可以用恰当的数学式来表达。

费歇尔之后的统计学家又提出了其他的准则，费歇尔自己也在后来的论文中提出了一些次要准则。剔除所有这些准则中的混乱不清的东西之后，剩下的最重要的元素就是，应该把统计量本身视为随机的，而好的统计量一定有好的概率特性。对于某一特定数据集，我们永远不知道一个统计量的值是否正确，只能说我们用一种方法得出来一个符合这些准则的统计量。

在费歇尔提出的三项基本准则中，"无偏性"准则最引入关注，这或许是由于"偏误"（bias）这个词带有某种贬义。一个有偏的（biased）统计量似乎是谁都不想要的某个东西。美国食品和药物管理局的正式指导准则就提出警告，要大家使用"避免有偏"的方法。有一种非常奇怪的分析方法（将在第 27 章里详细讨论），叫做"意向治疗"（intent to treat），已经成为占优势的医学试验法，因为，这种方法仍能保证结果是无偏的，尽管它忽略了有效性的准则。

事实上，一些有偏的统计量的应用常常极为有效。据费歇尔的研究，用来确定净化城市供水系统中氯浓度的标准方法，依据的就是一个有偏（但满足一致性与有效性）的统计量。所有这一切也是科学社会学 9the sociology of science）中的一类研究课题——为准确定义一个概念而创造出来的一个词，怎样将情感好恶的包袱也带到了科学中来，并对人们的行为产生了影响。

费歇尔的极大似然法

当费歇尔研究了这些数学问题之后，他认识到，用 K·皮尔逊的方法来计算分布参数所生成的统计量未必是一致的，而且经常是有偏的，他也认识到还存在着更加有效的统计量可以利用。为了得到一致且有效（但未必无偏）的统计量，费歇尔提出了被他称之为"极大似然估计量"（maximum likelihood estimator, MLE）的一个概念。

随后，费歇尔证明了 MLE 总是一致的，而且证明了如果人们认可几个被认为是"正则性条件"（regularity conditions）的假定，那么 MLE 是所有统计量中最有效的。此外，费歇尔还证明了，即便 MLE 是有偏的，也可以计算出其偏差的大小，然后将其从 MLE 的估计值中减掉，从而得到一个一致、有效且无偏的修正统计量[10]。

费歇尔的似然函数（likelihood function）席卷了整个数理统计学界，迅速成为估计参数的主要方法。极大似然估计只存在一个问题，就是在试图求解 MLE 时所涉及的数学问题，其难以对付的程度确实令人望而生畏。费歇尔的论文里写满了一行又一行的复杂代数式，用来说明不同分布的 MLE 数学公式的推导过程。他的方差分析和协方差分析的运算法则显示出他极高的数学造诣，去处过程中他设法在多维空间里利用巧妙的代入与变换，导出最终为使

[10] 在 20 世纪 50 年代，来自印度的 C·R·拉奥（C.R. Rao）与任教于美国霍华德大学（Howard University）的大卫·布莱克韦尔（David Blackwell）指出，就算费歇尔的正则性条件不成立，仍然有可能由 MLE 构造出一个最有效的统计量。这两个人分别独立地得出相同的定理，因此，以发现者命名了拉奥－布莱克韦尔定理（Rao-Blackwell theorem），两位发现者也确实因此而享誉学术界，成为施蒂格勒误称定律的一个例外。

用者所需要的 MLE 的计算公式。

尽管费歇尔具有非凡的独创性，但在多数情况下，对于 MLE 的潜在使用者来说，仍然难以驾驭所必需的高深数学知识。20 世纪后半叶的统计学文献中有许多非常睿智的文章，它们运用简化的数学方法，在某些实例中得到了相当理想的 MLE 的近似值。在我自己的博士学位论文里（大约写于 1966 年），我只能将就着不得不接受这样一个事实，即只有在能够得到非常多的数据时，我的问题的解才是好的。假定我有大量的数据，就能把似然函数简化到可以计算出挖 MLE 值的程度。

后来出现了电脑。电脑并非人脑的竞争对手，电脑只是一个巨大而有耐力的数字处理设备。它从不会厌烦，从不会困倦，也不会犯错误。它一而再、再而三地重复着做那些同样繁琐的计算，数百万次地一再重复。用所谓的“迭代算法”（iterative algorithms），它能算出 MLE 值。

迭代算法

最早的一种迭代数学方法好像出现在文艺复兴时期（虽然数学史学家大卫·史密斯（David Smith）在他 1923 年出版的《数学史》（History of Mathematics）中声称，早在古埃及和中国的文字记载中就已经发现了这种方法的实例）。当资本主义曙光初露之时，在意大利北部刚刚建立起来的商业银行或商号中就碰到一个基本问题：每个小小的城邦或国家都有自己的倾向，所以商号必须能算出如何在各倾向之间兑换；比如说，如果汇率是雅典钱币 14 德拉克马（Athenian drachma）换一个威尼斯币达克特（Venetian ducat），那么用威尼斯的 127 达克特买来的一堆木材，价值多少雅典的德拉克马呢？如今，我们有能力用代数符号来解答这个问题。还记得高中的代数吗？若 X 等于雅典德拉克马的值，则……

尽管当时的数学家已经开始发展代数学，这种简单的计算方法仍不能为大多数人所用。银行家用的是一种叫做“试位法”（rule of false position）的计算方法。由于每家商号都确信自己的换算规则是“最好的”，所以每家商号都有自己的店员。罗伯特·雷科德（Robert Recorde，1510-1558），这位 16 世纪的英国数学家，在普及代数符号上功绩卓著。为了把代数的威力与试位法则相对照，他在 1542 年写了一本书“The Grovnd of Artes”，书中说明了试位法：

Gesse at this woorke as happe doth leade.
By chaunce to truthe you man procede.
And firste woorde by the question,
Although no truthe therein be don.
Suche falsehode is so good a grounde.
That truthe by it will soone be founde.
From many bate to many more,
From to fewe take to fewe also.
With to much ioyne to fewe againe,
To to fewe adde to manye plaine.
In crosswaied multiplye contrary kinde,
All truthe by falsehode for to fynde.

雷科德的这篇 16 世纪的英文说的是：你先猜一个答案，并把它代入问题中，由此你会得到一个结果，而它和你想要的结果之间会有些差异。有了这个差异，接着你可以用它再产生一个更好的猜测，再用这个新的猜测得到一个新的差异，这个差异又会产生出另一个新的猜测值。如果在计算这个差异的过程中，你做得足够聪明，这一连串的猜测值会最终接近正

确的答案。对试位法来说，只要迭代计算一次，第二次猜测通常总能得到正确答案；而费歇尔的极大似然估计法，可能要迭代数千次甚至数百万次才能得到一个理想的答案。

然而，对一台任劳任怨的电脑，区区几百万次的迭代又算得了什么呢？在当今世界，不过是一眨眼的工夫。但在不久前，电脑的功能还不够强大，速度也很慢。在 60 年代末，我有个可以编写程序的台式计算机，是一种可以做加、减、乘、除的原始的电子工具。不过它还有个容易很小的内存，可以放进去一个程序，让它完成一系列的自述去处。这些运行的功能之一还能改写程序，因此，可以在这台可编程的计算机上运行迭代计算，只是要花很长的时间罢了。一天下午，我编好了计算机程序，检查了前几个步骤，确信我写的程序准确无误，然后，关掉办公室的灯就回家了。与此同时，这个编好了程序的计算机就开始了加减乘除的去处，静静地从它的电子结构内部发出喃喃的低语语，而且每隔一会儿就会按程序打印出一个计算结果。连接在计算机上的打印机是一个噪音很大的压缩设备，打印的时候会发出很响的“卡嗒、卡嗒”的声音。

那天晚上，保洁员到办公楼里清扫，其中一个人带着扫帚与废纸篓走进我的办公室。黑暗中，他听到了一种“嗡嗡嗡”的声音，他能看见在一遍又一遍进行加减的计算机上有只眼睛发出忽明忽暗的蓝光。突然，机器醒了过来，“卡”地响了一声，接着又“卡、卡、卡……卡嗒、卡嗒、卡嗒、”地响起来。后来他告诉我，那可真是一次让他毛骨悚然的经历。因此他要求我，如果下次计算机正在运行时，让我一定在办公室门口留一个提示纸条通知他们。

今天的电脑运行快得多了，甚至可以分析更加复杂的似然函数。哈佛大学的纳恩·莱尔德（Nan Laird）和詹姆斯·韦尔（James Ware）教授发明了一种异常灵活、功能异常强大、叫做“EM 演算法”的迭代过程演算法。在我订阅的统计学期刊里，每一期新杂志都会介绍某人如何采用他或她的 EM 演算法解决了一度被认为无法解决的难题。另有一些算法，名字颇富想象力，像“模拟退火法”（simulated annealing）、“克利金法”（kriging）等等，也不时地出现在文献中；还有“大都会”（Metropolis）算法或“侯爵”（Marquardt）算法，以及其他一些以发明者自己命名的算法。有一些很复杂的软件包，用成千上万行的程序编码，使这些迭代运算以“用户界面友好”的特点变得易于操作。

费歇尔的统计估计方法大获全胜，极大似然法统计了世界，而 K·皮尔逊的方法则被尘封在被遗忘的历史角落里。然而，就在这个时候，20 世纪 30 年代，当时费歇尔对数理统计理论的贡献终于得到了承认，他 40 多岁并且正值其事业鼎盛时期，就在那一刻，出现了一位名叫奈曼的年轻的波兰数学家，他对费歇尔一味遮掩却并没有真正解决的某些问题提出了质疑。

第 8 章　致死的剂量

每年的 3 月，生物统计学会都要在美国的南部城市召开一次春季会议，我们这些在北部生活和工作的人就借此机会南下，到路易斯维尔（Louisville）、孟斐斯（Memmphis）、亚特兰大（Atlanta）或新奥尔良（New Orleans），在会议结束后回家前的几周，去呼吸春天的清新空气，观赏原野中盛开的鲜花和果园里花繁叶茂的果树。同其他的科学会议一样，会议期间会有三到五位论文作者在会上口头宣读他们的论文，然后与会者与演讲人就论文的内容展开热烈的讨论，询问某些思想的出版，或提出其他可以替代的方法。通常，上午的会议分成两个分会场同时进行。最后的会议一般在下午 5 点前后结束，与会者回到宾馆各自的房间。一个小时或一个半小时之后他们又会分头聚在一起，相约着出去找一家喜欢的餐馆共进晚餐。

开会的当天，一般人总能在会场上遇到一些朋友，并绝好了会后一同去吃晚饭。但是有一天我却错过了约人就餐的时机。我和那天下午的一位论文演讲者进行了一场长时间的且饶有兴趣的讨论，他是当地人，散会后可以直接回家，因此我没有邀他一起吃饭。我们的谈话结束的时候，大厅里已经空荡荡的，人都走光了。我联系不上任何人，就回到房间给太太打电话，与孩子们在电话上聊了几句，随后就下楼到宾馆的前大厅，心想说不定会碰上一伙我认识的人，可以和他们一道活动。

但是，大厅里几乎空无一人，只有一个身材高大的白头发男人，他独自坐在一张罩着椅套的椅子上。我认出他是切斯特·布利斯（Chester Bliss），我知道他发明了一些基本的统计模型。那天上午在我参加的那个分会场，他还宣读了一篇论文。我朝他走过去，做了自我介绍，并称赞他上午的发言。他邀请我坐下，我们就坐在那里聊了一阵子统计与数学。不错，我们的确是在聊着这样的话题，我们甚至可以用这个话题来开玩笑。显而易见，我们俩谁也没有晚餐的约会，于是我们决定一起去吃晚饭。他可真是个令人愉悦的就餐伙伴。那天的晚餐，我听他讲述了自己丰富的阅历。以后的几年，我们常在开会的时候碰面，有时还会相约一同用餐。他在耶鲁大学的统计系任教，所以，每当我参加由耶鲁大学统计系主办的研讨会时，就经常能见到他。

布利斯出身于美国中西部一个殷实而融洽的中产阶级家庭，父亲是医生，母亲掌管家务，有几个兄弟姐妹。他起初对生物学感兴趣，念大学时学的是昆虫学。20 世纪 20 年代末，他大学毕业后，以一个昆虫学家的身份供职于美国农业部，并且不久就参与 了研制杀虫剂的工作。很快，他认识到，在田间试验杀虫剂会受到许多无法控制变量的干扰，使结果难以解释，于是，他把昆虫带到实验室里，做了一系列的实验。这时，有人把费歇尔所写的《研究工作者的统计方法》一书介绍给他，以此为起点，他一边努力去领悟费歇尔在这本书中介绍的许多统计方法的深层次内涵，一边又阅读了费歇尔更多数学论文。

概率单位分析

在费歇尔统计方法的引导下，不久，布利斯说开始了他在实验室内的实验。他把昆虫分成几组，养在广口玻璃瓶里，然后用不同成分和不同剂量的杀虫剂来实验。在他做这些实验的过程中，发现了一个值得关注的现象：无论他配制的杀虫剂尝试有多高，在用药之后总会有一两只昆虫还活着；此外，无论他怎么稀释杀虫剂，即便只是用了装过杀虫剂的容器，试验结果也总会有几只昆虫死掉。

有了这些显著的变异，如果能依据皮尔逊的统计分布建立一个数学模型来分析杀虫剂的作用，这将是非常有用的。但是如何建立这个模型呢？你很可能会回想起高中代数课上，当书本翻到解文字题时那令人头疼的时刻：A 先生和 B 先生共同在静止的水中划船；或者在平

稳流动的水中逆流而上；或者他们会把油与水混在一起；或者让他们来来回回地运球。无论哪一种问题，这种文字应用题总是给出一些数字，然后问一个问题，可怜的学生就必须把这些文字转换为数学公式，并解出未知数 x。你或许能回想起当初是如何哗哗地翻查着教科书，拼命地寻找一个类似的并且已经解出答案的例题，然后把文字应用题的新数字塞进这道例题所用的公式中去。对高中的代数课而言，总有人已经把相关问题的数学公式列了出来，要么老师知道这些数学公式，要么能在与教科书配套的教师手册里找到这些公式。然而，试想有这样一个文字应用题，没有人知道如何将它转化为数学公式，没有人知道问题当中哪些数据是多余的，哪些应该是没用的，而一些至关重要的信息又常常缺失，况且教科书中也没有事先已经解出来的类似例题。这就是当你设法把统计模型应用到现实生活中去的时候所面临的情景，这也正是当布利斯打算采用概率分布这种新的数学思想来分析他的杀虫剂实验时所遭遇的困境。

为此，布利斯发明了一种他称之为“概率单位分析”（probit analysis）的方法，这项发明需要一种非凡跨越的原创性思想。这种方法中的任何思想，甚至哪怕是应该如何去做的启示，都未曾出现在费歇尔的“学生”的、亦或其他什么人的著作中。他之所以使用“概率单位”（probit）这个词，是因为他的模型建立了“杀虫剂的剂量”与“使用该剂量时一只虫子会死掉的概率”这两者间的关系。他的模型中生成的最重要的参数谓之“半数致死剂量”（50 percent lethal does），通常用“LD-50”来表示，是指杀虫剂能以 50%的概率杀死虫子的剂量。或者说，如果施用这种杀虫剂来对付大量的虫子，那么用“LD-50”的剂量，将有 50%的虫子被杀死。布利斯模型的另一个推论则是：对一只特定的用做实验标本的虫子，要确定杀死它所需要的剂量是不可能的。

布利斯的概率单位分析已被成功地应用于毒物学（toxicology）。从某种意义上说，源于概率单位分析的认识已经形成了毒物学这门科学的主要基础。16 世纪的医师 P·A·帕拉赛瑟斯（P. A. Paracelsus, 1493-1541）有一句名言：使用过量，什么都是毒药。概率单位分析为帕拉赛瑟斯首创的这个信条奠定了数学基础。按照帕拉赛瑟斯的这个信条，只要剂量足够大，任何东西都可能成为毒药；而只要剂量足够小，任何东西都是无害的。而布利斯则为了这个信条增加了与那些个案结果联系在一起的不确定性。

之所以会有那么多愚蠢的吸毒者，在古柯碱、海洛因或安非他命的作用下，或已毙命于街头，或变得极度虚弱，原因之一就在于，他们看到其他人同样服用这些毒品却没有死于中毒。他们就如同布利斯实验用的那些虫子，环顾四周，看到有些同伴依然活着。然而，即使知道某些个体还活着，也无法确定一个给定个体能否幸免于一死。我们根本没有任何办法能够预见某一独特个体对药物剂量的反应。就像皮尔逊统计模型里的那些个别观测值一样，它们都不是科学研究所关注的“事件”。惟有那些抽象的概率分布及其参数（如 LD-50，半数致死剂量）才是能够估计的。

布利斯的概率单位分析一经提出[11]，其他研究人员也跟着提出了各种不同的数学分布。

---

[11] 施蒂格勒误称定律在概率单位分析里得到了印证。显然，布利斯是第一个提出这种分析方法的人。但是，这个方法要用到一个复杂的计算表中的两阶段迭代计算和内插法。1953 年，美国氰胺公司（Cvanamid）的弗兰克·威尔科克森（Frank Wilcoxon）做出一组图表，用户只要用直尺去对照一套标好的直线，就能计算出概率单位的值。后来，这个方法由 J·T·利奇菲尔德（J. T. Lichfield）和威尔科克森联名发表在一篇论文中。为了证明这种绘图解法能得出正确答案，两位作者在论文的附录中列出了由布利斯与费歇尔提出的数学公式。20 世纪 60 年代末的某个时候，有修水知名的药理学家把这篇论文拿给了一位不知名的程序员，这个程序员就用这个附录写成了一个可进行概率单位分析的计算机程序（用布利斯的迭代算法），那个程序的文件记录引用了利奇菲尔德和威尔科克森的论文作为参考文献。不久，其他概率单位分析的计算机程序开始出现在各大公司和各大学的药学系，而所有这些都源于最初的这个计算机程序，因此，也都在他们的文件记录中把利奇菲尔德和威尔科克森的论文当作参考文献。最终，用这些计算机程序所做的概

现代用来计算“LD－50”半数致死剂量的计算机程序，通常都会提供几种不同的模型让用户选择，这些模型都是在布利斯的原创基础上经过改进之后提出来的。用实际数据所做的研究表明，尽管在估计非常低的概率时，如“LD-10”，由这些不同模型得出的估计值是有差别的，但在“LD-50”上的估计值都非常接近。

我们完全可以运用概率单位分析或选择其他模型来分别估计一个不同的致死剂量，如“LD-25”或“LD－80”（25%的死亡剂量，或 80%的死亡剂量）。不过，离 50%点越远，就越需要更大规模的实验才能得到理想的估计值。我自己就曾参与过一项研究，要确定某种能在老鼠身上致癌的化合物的 LD-01（1%的致死剂量）是多少。我们的实验用了 65000 只老鼠，最终的分析结果表明，我们还是没能得到使 1%老鼠致癌的化合物剂量的理想估计值。依据那项研究的数据推算，要想得到一个可接受的 LD-01 的估计值，我们得需要几亿只老鼠！

布利斯在列宁格勒

C·布利斯在概率单位分析上的开创性研究，到 1933 年却被迫中断了。那年，弗兰克林·D·罗斯福（Franklin D. Roosevelt）当选为美国总统。在竞选总统期间，罗斯福明确声称是联邦政府的赤字导致了经济萧条，并且保证他当选后会消减政府赤字，缩小政府部门的规模。虽然这并不是“新政”（the New Deal）最终的行为，却是竞选的诺言，因此这位新总统就职之后，他的一些内阁成员就遵照总统的竞选诺言，开始解雇一些非必要的政府工作人员。

那位协助农业部副部长负责研制新式杀虫剂工作的助理，当他在视察这个部门所做的工作时，发现有人居然不到有虫子的田间去做实验，反而无聊地躲在实验室里不厌其烦地用杀虫剂来做实验。于是，布利斯的实验室被关闭了，布利斯也被解雇了。当时正值严重的大萧条时期，他发现自己根本找不到工作。尽管布利斯曾发明了概率单位分析，但对于一个失业的昆虫学家，特别是一个与实验室的昆虫，而不是野外的昆虫打交道的昆虫学家来说，找不到工作实在是不足为奇。

布利斯与费歇尔取得了联系。费歇尔刚刚在伦敦得到一个新职位，他答案举荐布利斯，并给他一些实验设备，不过他不能给他一个工作岗位，因此也没有办法付给这位美国昆虫学家工作报酬。尽管如此，布利斯还是不得不去了英国。他与费歇尔及其家人一起住了几个月，并与费歇尔一起协作进一步完善了概率单位分析的方法论。费歇尔在布利斯的数学去处中发现了几处错误，并提出修改建议，得到的最终统计结果更为有效。布利斯按照费歇尔的修改建议，发表了一篇新论文。而费歇尔也把那个必不可少的统计表，补充编到他自己与弗兰克·耶茨（Frank Yates）联名写的有关统计表的那本书的新版中去。

布利斯在英国住了不到一年，费歇尔就为他找到了一份新工作，是在苏联的列宁格勒植物研究所（Leningrad Plant Institute）。试想一下，这个来自美国中西部地区中产阶级家庭、对政治漠不关心、而且永远不会学第二种语言的又高又瘦的家伙，随身带着只装了几件换洗衣服的一个小行李箱，乘火车只身穿越欧洲大陆，终于到达列宁格勒火车站时的情景。而那时的俄国恰逢斯大林领导下的大清洗运动。

布利斯到达列宁格勒之后不久，聘请他来苏联的那个人的老板就被召到莫斯科去了，而

率单位分析开始出现在药理学和毒物学的文献中，当然，利奇菲尔德和威尔科克森的论文也被当作概率单位分析的“源头”列入参考文献中。这样，在对大多数已发表的科研论文中被引用的参考文献进行列表统计的《科学引文索引》（Science Citation Index）里，利奇菲尔德和威尔科克森的这篇论文已经成为历史上被引用次数最多的论文之一——倒不是因为他们做了什么了不起的大事，而是因为布利斯的概率单位分析已被证明是非常有用的工具。

且从此销声匿迹。一个月之后，那个聘请布利斯来苏联的人也被召到莫斯科去了，而且在返回途中“畏罪自杀”。负责布利斯旁边那个实验室的主管，也在某一天仓惶弃职，穿过拉脱维亚边境逃出了苏联。

就在这个时候，布利斯认真着手展开他的实验工作。他选了几组俄罗斯本地的害虫，用各种不同化合成分的杀虫剂来对这几组害虫进行试验，算出其概率单位极其“LD-50”半数致死剂量。他在研究所附近的房子里租了一个房间，他的俄罗斯女房东只会说俄语，而布利斯只会说英语。不过他告诉我，用各种手势加上亲切的微笑，他们相处得相当融洽。后来，布利斯遇见了一个来自美国的年轻女人，她为了投身于俄国伟大的共产主义实践，中断大学学业，满怀着年轻人的理想主义和马克思列宁主义的盲目崇拜来到苏联。她把可怜的只会说英语的布利斯当成好朋友，帮他购物、熟悉环境。此外，她还是当地的一个共产党员。党组织对布利斯的一切了如指掌，他们知道他何时受聘，何时抵达俄国，住在什么地方，以及在实验室里都做了些什么。

有一天，那女孩告诉他，党员里有些人已经认定他是美国间谍。她竭力为布利斯辩护，向他们解释布利斯是个单纯而又天真的科学家，只热衷于自己的实验。但是这些猜疑已经通报给了莫斯科当局，他们已经派出了一个委员会到列宁格勒来进行调查。

调查委员会就在列宁格勒植物研究所召开审查会，把布利斯叫来面对他们接受审问。当他走进审问室的时候，已经知道调查委员会里每个人的身份了，当然是他的女朋友透露的。他们几乎还没来得及调查完最初的几个问题，就在这时，布利斯对他们说：“我看到某某教授也坐在你们中间（告诉我这段往事的时候，布利斯已经不记得这位教授的姓名了），我一直在读他的论文。请告诉我，他提出的这种农业试验方法，是遵照圣人马克思和圣人列宁的绝对真理吗？“翻译踌躇着吞吞吐吐地把它这句话译了出来，刚一译完，审查委员会的委员们便一阵忙乱，他们要求布利斯对此做进一步的阐述。

“某某教授的方法”，布利斯接着又问：“就是正规的党的方式吗？就是按照党所要求的做法进行的农业试验吗？”

最终委员会给他的答案是，没错，这确实是做事情的正确方法。

于是布利斯说：“如果是那样的话，我就是违背了你们的信仰。”接着他进一步解释，如果按照这个教授提出来的做法，农业试验研究必须用很大面积的土地，而且所有这些农田都得用同样的实验方式来处理。布利斯说，他认为这样的试验是无益的，并且阐明他一直在倡导的方法，就是把农田分成很多小块地，以不同的方式处理相邻的地块。

审查工作没有再深入进行下去就结束了。那天傍晚，他的朋友告诉他，委员会已断定他不是间谍。他们认为他太率真了，透明得一眼就可以看穿，或许真是如她所说，他是一个头脑单纯、只关心他的实验的科学家。

其后，布利斯在列宁格勒植物研究所工作了几个月。他再也没有任何顶头上司了，他自己认为怎么做最好就怎么做。但是，他必须加入由实验室工作人员组成的工会组织，当时，每个在俄国工作的人都必须加入某个由政府控制的工会组织。除了这一点规定之外，他们就不管他了。在 20 世纪 50 年代，美国国务院还曾因为他一度属于一个共产党的组织，而拒绝给他签发美国护照。

突然有一天下午，他的女朋友冲进实验室，告诉他：“你必须马上离开。”他坚持说他的实验还没有做完，实验结果还没有详细记录下来，坚持要做完这些才肯离开。女友把布利斯从实验报告堆中拽出来，逼他赶紧穿上外套，告诉他刻不容缓，必须丢弃所有的一切，必须马上离开。刀子守候着催促着他，看着他装好那个小小的提箱，告别了女房东。女友把他送到火车站，临行前坚持要他在安全抵达里加（Riga，现拉脱维亚共和国的首都）时给她打个电话。

到了 20 世纪 60 年代，苏联的政治局势有了些微的松动，苏联的科学家重新回到国际科

学团体中来。国际统计学会（International Statistical Institute, C·布利斯曾是该学会的会员）在列宁格勒召开了一次国际会议，会议期间，布利斯抽空去探访那些 30 年代的老朋友，但他们都已故去。他们当中，有的是在大清洗时期被杀，有的死于第二次世界大战，只有他当年的女房东还活着。见面时，他们不停地用各种手势，不断地点头，互致问候，并亲切拥抱，布利斯用英语低声地表达着对她的美好祝福，她则以俄语回应。

第 9 章　钟形曲线

读完这本书的前八章，你也许会以为统计革命只是发生在英国。从某种意义上说，这倒也是事实，因为最先将统计模型应用于生物研究和农业研究的，的确是在英国，还有丹麦。在费歇尔的影响下，统计学方法很快就传到了美国、印度、澳大利亚和加拿大。正当统计模型的实际应用在说英语的国家和地区推广之际，由于欧洲大陆长期形成的一种数学传统，使得欧洲的数学家正在研究与统计建模有关的理论问题。

这些理论问题中，最为重要的是中心极限定理（central limit theorem）。直到 20 世纪 30 年代初，这还是个未经证明的定理，或者说只是一个猜想（conjecture），因为许多人都信其为真，却没有一个人能证明它成立。费歇尔早在研究似然函数值的理论时，就曾假设这个定理是成立的；而回溯到 19 世纪初，法国数学家皮埃尔·西蒙·拉普拉斯也用这个推论证明了他的最小平方法（method of least squares）。此外，心理学这门新兴科学也是根据中心极限定理开创了智力测验技术与精神疾病量表。

什么是中心极限定理？

大量数据集合的平均数都有一个统计分布，而中心极限定理则阐明，无论初始数据是怎么来的，这个分布都可以用正态概率分布来逼近。这个正态概率分布与拉普拉斯的误差函数（Laplace's error function）相同，有时也叫做高斯分布（Gaussian distribution），而在浅显通俗的普及书里，也常被称为"钟形曲线"（bell-shaped curve）。在 18 世纪晚期，亚伯拉罕 •棣莫弗（Abraham de Moivre）已经证明，由机会博弈(games of chance)所得数字的简单集合符合中心极限定理。然而，在此之后的 150 年里，对这个猜想的证明没有丝毫的深入进展。

用正态分布来描述大部分数据都是正确有效的，因此，中心极限定理普遍被认为是一个正确的猜想。一旦假定数据服从正态分布，数学上的处理就容易多了。正态分布具备某些非常优良的性质：如果有两个随机变量服从正态分布，那么两变量之和也同样服从正态分布。就一般而言，正态变量的各种类型的和与差也都服从正态分布。因此，由正态随机变量（variate）推演得出的许多统计量，其自身也服从正态分布。

正态分布只有 K·皮尔逊四个参数中的两个——平均数和标准差，另外两个参数对称性偏度（symmetry）和峰度(kurtosis)均为零。因此，一量知道了平均数和标准差这两个参数值，其他的一切也就一清二楚了。费歇尔曾指出，由一组数据得出的平均数与标准差的估计值就是他所说的充分估计量（sufficient estimator），因为这两个参数值已经把这些数据中所有的信息都包括在内了。既然这两个参数值已经涵盖了能够从那些原始测量值中揭示出的一切，就根本没有必要去占有任何原始测量值了。如果有足够的测量值可以用来相当精确地估计出平均数与标准差，就不再需要其他任何测量值了，任何为搜集这些数据所做的努力，都不过是浪费时间而已。例如，有两个重要指标服从正态分布，如果你正打算得出这样一个正态分布的那两个参数，那么你只需要收集大约 50 个测量值就足够了。

正态分布的这种数学上便于处理的特性，使科学家能够构建一个复杂关系模型。只要其基本分布是正态的，费歇尔的似然函数通常就有了以简单代数进行处理的一种形式。即便模型复杂到必须用迭代运算法去解的程度，只要其分布是正态的，用纳恩·莱尔德（Nan Laird）和詹姆斯·韦尔(James Ware)的 EM 演算法去解，就变得轻而易举了。由于正态分布在数学上的计算处理非常敏捷，因此在建模时，统计学家常常要假定所有的数据都服从正态分布。不过，做这样的假定就不能不援引中心极限定理。

但是，中心极限定理是否成立？说得更准确一点，它在什么条件下成立？

在 20 世纪 20 年代和 30 年代，斯堪的纳维亚地区、德国、法国和苏联的一批数学家，

运用 20 世纪早期发明的一套新的数学工具，倾心于上述这些问题的研究。但就达个时候，整个人类文明都正面临着一场日益迫近的浩劫——那些极权主义的国家的恶性膨胀。

数学家并不有昂贵设备的实验室。在 20 世纪二三十年代，黑板和粉笔就是一个数学家最具代表性的实验设备。对数学研究而言，用黑板比用纸张更方便，因为数学研究过程的演算总免不了出错，而黑板上的粉笔字很容易擦掉。几乎没有数学家是关起门独自做研究的，只要你是一个数学家，你就必定要同其他的数学家一起讨论自己在研究的问题，你就必定要接受别人对你那些新想法的批评审视。在数学研究过程中太容易出错，或者太容易在研究中隐含着自己毫无察觉而在别人看来却是显而易见的假设。有一个数学家的国际组织，在这个团体中，数学家们书信往来、开会、审阅彼此的论文，经常交换相互的批评和质疑，探究分歧所在。20 世纪 30 年代初期，德国的威廉·费勒（William Feller）和里夏德·冯·米泽斯（Richard von Mises），法国的保罗·利维(Paul Lévy)，俄罗斯的安德烈·柯尔莫哥洛夫（Andrei Kolmogorov），斯堪的纳维亚的亚尔·瓦尔德马·林德伯格（Jarl Waldemar Lindeberg）和哈拉尔德·克拉美（Harald Cramer），奥地利的亚伯拉罕·沃尔德（Abraham Wald）和埃尔门·哈特利（Herman Hartley），意大利的圭多·卡斯泰尔诺沃（Guido Castelnuovo），还有许多其他数学家也都在这个团体中，其中不乏那些利用新工具来检验中心极限定理这个猜想的数学家。

然而，这种自由轻松、无拘无束的相互交流不久就将不复存在。它将毁于斯大林的肃反运动、纳粹的种族灭绝和墨索里尼的帝国梦。黑暗笼罩着欧洲。斯大林正把非法操纵的示众式的公开审讯同半夜里的秘密逮捕结合运用到了极致，处决、恐吓和威胁任何一个受到他偏执狂式的无端猜疑的人。起初，希特勒及其罪大恶极有党羽把犹太裔教授从各大学里清洗出去，随后将他们关进惨无人道的集中营。墨索里尼则把国人强行禁锢在他所谓的“组合国”（Corporate state）所划定的各个社会等级中。

“死亡万岁！”

这一猖獗的、反理智主义（anti-intellectualism）的极端事件，就发生在西班牙内战时期。当时长枪党的党徒们（以西班牙的法西斯主义者闻名）已经占领了古老的沙拉曼卡大学（University of Salamanca）。该大学的校长是享誉世界的西班牙哲学家米格尔·德·乌纳穆诺（Miguel de Unamuno），当时他已经 70 岁出头了。长枪党的米连·阿斯特赖（Millan Astray）将军，一个在先前的战争中失去了一条腿、一只手臂和一只眼睛的残疾人，当时任这个刚以武力控制了西班牙的恶势力的宣传部长。他的座右铭就是：“死亡万岁！”如同莎士比亚笔下的国王理查德三世，阿斯特赖身体上的残缺不全恰恰映射出他扭曲的邪恶心灵。有一次，长枪党在沙拉曼卡大学的纪念大厅举行盛大的庆典，台上有新指派的省长、弗朗西斯科·佛朗哥（Francisco Franco）夫人、M·阿斯特赖、沙拉曼卡的大主教，还有年事已高的乌纳穆诺，他是被当作被征服的战利品拖到台上的。

“死亡万岁！”阿斯特赖高声狂呼，挤满了人群的大厅里随声附和着他的喊叫。又有人高呼：“西班牙！”大厅里的人也跟着喊。“西班牙！死亡万岁！”穿着蓝色制服的长枪党的党徒们齐刷刷地站起来高呼，并朝着台上的佛朗哥肖像行法西斯的举手礼。就在这一浪高过一浪的叫嚣声中，乌纳穆诺站起身来，从容地走向讲台，镇静地开始演讲：

> 你们大家都记住我的话。你们都了解我，并且知道我不可能保持沉默，因为沉默也可以解释为默认，沉默中常常意味着谎言。我想对刚才的演讲做个评论，我们不防就叫它“M·阿斯特赖将军的演讲”吧……。就在刚才，我听见一种嗜尸成癖的愚蠢无知的叫嚣：“死亡万岁！”而我，一个终生致力于各种悖论研究的人……我必须告诉你们，作为一个权威，这种荒诞怪异、语无伦次的谬论让我恶心。阿斯特赖将军是个残疾人……

他是战争造成的一个残疾人……。不幸的是，眼下的西班牙这种残疾太多了。而且不久，如果上帝不能拯救我们，这种残疾人甚至还会更多……。

M·阿斯特赖把乌纳穆诺推到边上，厉声吼叫："该死的臭知识分子！死亡万岁！"与他的叫嚣相呼应，那些长枪党徒们蜂拥而上，抓住乌纳穆诺。但是，老校长仍然继续说道：

这里是知识的殿堂，而我才是这个殿堂的领袖。是你们亵渎了这个神圣的地方。你们可以凭借极其残暴的兽行获胜，但是你们无法得到人们的认可。因为要让人认可必须靠说服而不是压服，要达到说服的目的所必须具备的东西，恰恰是你们所没有的，那就是理智和正义……。

乌纳穆诺遭到软禁，不出一个月就被宣告"自然死亡"。

苏联的大清洗运动切断俄国数学家与欧洲其他地方的联系；希特勒的种族政策几乎毁掉了德国的大学，因为欧洲许多伟大的数学家要么是犹太人，要么是与犹太人联姻，而非犹太裔的那些数学家又大多是反纳粹的。结果，威廉·费勒去了美国的普林斯顿大学（Princeton University），亚伯拉罕·沃尔德到哥伦比亚大学（Columbia University）去了，埃尔门·哈特利和里夏德·冯·米泽斯去了英国伦敦，埃米尔·J·冈贝尔逃到了法国，埃米纳脱（Emmy Noether）在美国国宾夕法尼亚的布林莫尔学院（Bryn Mawr College）求得一个临时工作。但是，并非每个人都逃得脱。不能出示证明受聘到美国去工作的那些人，美国移民局对他们总是大门紧闭；而拉丁美洲那些国家的国门则由于那些小官僚的反复无常而时开时关。纳粹军队占领了波兰首都华沙后，大肆搜捕能找到的华沙大学所有的教授和学者，逮捕他们并惨绝人寰地将他们杀害，然后一起埋在一个巨大的坟墓里。在纳粹的种族主义世界里，波兰人和其他斯拉夫人只配做他们这些亚利安（Aryan）主人的奴隶，没有受教育的权利。欧洲那些历史悠久的大学里许多有培养前途的青年学生就这样被毁掉了。在苏联，大部分数学家都躲进了纯数学中去寻求庇护，而不敢在应用领域中做任何尝试。因为，那些从事应用研究的科学家，正受到斯大林令人不寒而栗的无端猜疑。

不过，在这些黑暗没有完全成为现实之前，欧洲的数学家们就已经解决了中心极限定理的证明问题。芬兰的亚尔·瓦尔德马·林德伯格和法国的保罗·利维分别发现了能够使中心极限定理这个猜想成立所必需的一组重叠的条件。这证明了至少存在三种解这个问题的方法，而且证明了中心极限定理不是只有一个单个的定理，而是有一组定理，其中每个中心极限定理都能从略有区别的一组条件中推导出来。到了 1934 年，中心极限定理（组）终于不再是猜想了，一个科学家必须要做的只是要证明林德伯格·利维条件（Lindeberg-Lévy Conditions）成立，那么中心极限定理就成立，于是，他就可以随意地把正态分布设为一个合适的模型。

林德伯格·利维条件与 U 统计量

然而，就一个特定情况而言，要证明林德伯格·利维条件成立很难。但在理解林德伯格·利维条件上倒有几分安慰，因为他们描述的条件看上去是合理的，而且在大多数情况下都是成立的。不过要证明其成立却是一个棘手的问题，这也正是战后远在北卡罗莱纳大学辛苦工作的瓦西里·霍夫丁（Wassily Hoeffding）在这个故事中竟会有如此重要地位的原因。1948 年，霍夫丁在《数理统计年报》（Annals of Mathematical Statistics）上发表了一篇论文，题目是"渐近正态分布的一组统计量"。

回想费歇尔曾把统计量（statistic）定义为：从观察到的测量值得出的、可用来估计其分布参数的一个数值。费歇尔还建立了有用的统计量应该具备的一些准则，在这个过程中，他还指出了利用皮尔逊的许多方法导出的统计量不符合这些准则。有很多种计算统计量的不同方法，其中的很多统计量都能满足费歇尔提出的准则。一旦计算出统计量，为了要用它，

我们必须知道它的分布。如果它服从正态分布，用起来就容易多了。霍夫丁提出了一种他所谓的“U-统计量（U-statistics），并指出一个统计量如果属于这种 U-统计量，则满足林德伯格·利维条件。正因为如此，我们只须指出一个新的统计量是否与霍夫丁的定义相一致，而不必去解那些很困难的数学来证明林德伯格·利维条件成立。霍夫丁所做的一切就是用一组数学必要条件取代另外一组。然而，霍夫丁的条件事实上很容易检查。因此，霍夫丁的论文发表之后，几乎所有的论文在证明一个新统计量服从正态分布的时候，都是通过证明该统计量是一个 U 统计量来完成的。

霍夫丁在柏林

第二次世界大战期间，霍夫丁处在一个不确定的微妙境况中。他 1914 年出生在芬兰，父亲是丹麦人，母亲是芬兰人。第一次世界大战之后，芬兰沦入俄罗斯帝国的统治，就在这个时候，霍夫丁随家人迁往丹麦，随后又迁往柏林，因此他拥有斯堪的纳维亚地区两个国家的双重国籍。1933 年他高中毕业，随后开始在柏林攻读数学。就在那个时候纳粹开始在德国掌权。预料到以后可能发生的事，霍夫丁就读的那所大学的数学系的系主任 R·冯·米泽斯早早地离开了德国，不久之后，为霍夫丁授课的其他许多教授，有的逃走了，有的被解除了职务。在动乱中，年轻的霍夫丁所选的课都是由一些低水平的教师来讲授的。即便如此，这些教师中的很多人也没能维持到把他们承担的课程教完，因为纳粹在持续不断地“净化”大学教师队伍，把大学教师中所有的犹太人和犹太人的同情者全都清除出去。

随同数学系里的其他学生一道，霍夫丁被迫去听路德维希·比贝尔巴赫（Ludwig Bieberbach）讲授的一堂课。比贝尔巴赫一直都是教师中的小角色，只是因为对纳粹党的狂热拥护，才合他成为数学系新的系主任。比贝尔巴赫这堂课讲的是“亚利安”数学与“非亚利安”数学的区别，他声称颓废的“非亚利安”（解读为犹太）数学家仰仗着复杂难解的代数符号做研究，相反，“亚利安”数学家则在更高贵、更纯粹的几何直觉领域里从事研究。结束了讲课的时候，他让学生提问题。坐在后排的一个学生问他，为什么偏偏是这个里夏德·库朗（Richard Courant，20 世纪初德国伟大的犹太裔数学家之一）运用几何洞察力创建了他的实分析理论（theories of real analysis）。此后，比贝尔巴赫再也没有就这个题目上过公开课。但是他创办了《德国数学》（Deutsche Mathematik）杂志，这个杂志很快就成为当政者眼中居第一位的数学期刊。

1940 年，霍夫丁完成了他的大学学业，像他这个年龄的其他男青年都要应征到部队去服兵役，但由于他的双重公民身份，并且当时的芬兰已成为德国的一个盟国这样的事实，他因此不必服兵役。他找到一份工作，在一家跨校际的精算杂志社的办公室兼职。与比贝尔巴赫创办的那个杂志不同，这是一个很难约到论文，因此也很难定期出版发生的杂志。霍夫丁甚至连寻找一份教书的工作都不能，因为他必须申请到正式的德国公民身份才有资格去教书。

1944 年德国政府宣布，具有“德国血统或相关血统“的非德国籍青年也要服兵役。不过，在霍夫丁体检的时候，发现他患有糖尿病而免于服兵役。这时他终于有了找工作的资格。他兼职的那家期刊的编辑哈拉尔德·格佩特（Harald Geppert）建议他从事某种军事应用方面的数学研究工作，他提这项建议的当时，期刊的另一个编辑赫尔曼·施密德（Hermann Schmid）也在场。霍夫丁犹豫了一下，然后，出于对格佩特的谨慎的依赖，他对格佩特说，任何一种与战争有关的工作都违背他的良心。施密德出身于一个普鲁士贵族家庭，霍夫丁希望他的家族荣誉感能让他对这次谈话守口如瓶。

随后的几天里，霍夫丁一直提心吊胆的，但什么事都没有发生，他得以继续他的研究。当俄国军队逼近柏林的时候，一天早上，格佩特在早餐里放了毒药喂给他年幼的儿子，随后

他和他的太太也服毒自杀了。1945 年 2 月，霍夫丁和他的母亲一起逃到汉诺威的一个小镇上，他们在那里的时候，这个地方成为英军占领区的一部分。而他父亲仍滞留在柏林，在那里，他被俄国秘密警察以间谍罪逮捕，因为他一度曾为美国驻丹麦的商务参赞工作过。好几年时间，他杳无音信，直到他设法越狱，又历尽千辛万苦逃到了西方。在此期间，年轻的霍夫丁于 1946 年秋天到达纽约，继续他的学业，后来应邀到北卡罗莱纳大学任教。

运筹学

纳粹的这种反理智主义、反犹太主义倒行逆施的结果之一，就是让第二次世界大战的同盟国因此而丰收了许多才华横溢的科学家与数学家，在他们的鼎立相助下打赢了这场战争。英国生物学家彼得·布莱克特（Peter Blackett）向海军部建议，武装部队应该请一些科学家来协助解决战略和战术上的问题。无论是哪个专业研究领域的科学家们，他们都训练有素，能够应用逻辑和数学模型来解决问题。他建议组成科学家的攻关小组，让这些小组从事有关战争问题的研究，由此诞生了一门新学科——“运筹学”（operational research，在美国称之为 operations research）。从事不同领域研究的科学家组成的科研小组联合起来共同研究，决定用远程轰炸机对付潜艇的最佳使用方案；为防空武器提供射击表；决定靠近前线的军火补给站的最佳选址；甚至还要解决军队的食物补给问题。

战争结束后，运筹学的应用由战场搬到了商场。那些在战争期间被征募到军队去服务的科学家已经证明了用数学模型和科学的思维来解决战事中的战术问题是多么有用。同样的步骤和许多相同的方法也能用来组织工厂里的生产，找出仓库与销售部门之间的最优关系，解决许多别的商务问题，均衡有限的资源，或改进生产与提高产量。从那时候起，大公司里大部分都设立了作业研究部门，而这个部门所从事的多数工作都与统计模型有关。

我在辉瑞公司工作的时候所做的几个项目，其目的都是为了改善对药物研究进行控制和提取新产品进行测试的方法，在所有这些研究中涉及到的一个重要方面就是，当条件可以满足时，有能力用正态分布去处理问题。

第 10 章 拟合优度检验

20 世纪 80 年代，出现了一种新型数学模型，激起了公众的遐想，主要是因为这种数学模型的名字——混沌理论（chaos theory）[12]。这个名字提示着某种形式的统计建模明显带有杂乱无序特征的随机性。创造了这个名字的人有故意避开使用随机（random）这个词的嫌疑。实际上混沌理论是尝试着在一个更高端的层次上，通过复兴决定论（determinism）来动摇统计革命。

回想一下，在统计革命之前科学所处理的那些“事件”，要么是已有的测量，要么是生成这些测量值的自然事件。伴随着统计革命，科学的事件变成了能左右测量值分布的参数。在早期的确定性方法中，有一个信条是，越精确的测量，对所考察的自然客体的描述也就越精确；而在统计方法中，分布参数有时候不必有一个自然客体，无论多么精确的测量系统，分布参数的估计值终究是有误差的。例如，在确定性方法中，重力常数是描述物体如何向地球下落的一个恒定不变的值；而在统计方法中，我们对重力常数的测量值永远都不会是一样的。为了“通晓”落体的性质，这些测量值分布的离散状态才是我们想要确立的。

1963 年，混沌理论专家爱德华·洛伦兹（Edward Lorenz）做了一个后来时常被引用的演讲，演讲题目为“巴西一只蝴蝶翅膀的翩翩舞动，会引起德克萨斯州的龙卷风吗？”洛伦兹的主要论点是，混沌的数学函数对初始条件非常敏感，初始条件的些微差异，经过多次迭代之后，中以导致全然不同的结果。洛伦兹相信，由于存在这种对初始条件微波差异的敏感性，以至于对所研究的问题不可能得出一个确定的答案。隐含在洛伦兹演讲中的是确定性假设，即理论上每一个初始条件都是促成某个最终结果的一个起因。这个被称之为“蝴蝶效应”（butterfly effect）的观念，已经被那些混沌理论的普及者们当作一个深邃而睿智的真理接受下来了。

然而，没有任何科学的证明揭示了这样一种因果关系的存在，也没有任何数学模型有准确的依据表明客观现实中存在着这一效应。它只是一种信念的表述而已，就其科学的有效性而言，它与关于鬼神的描述相去无几。而统计模型是用分布参数来对科学探索明确地进行解释，它们也是建立在对现实世界的一种信念所做的描述上。然而，我自己在科学研究上的经历让我确信，比起对信念的确定论的陈述，统计上的陈述更有可能是真实的。

混沌理论与拟合优度

混沌理论源于这样的观察：一个固定不变的确定性公式生成的数字有可能看上去是一个具有随机性的模型。早在一批数学家处理相对简单的迭代公式并绘出其结果的时候，就曾经发现过这种现象。在第 9 章，我曾经把一个迭代公式描述为：首先得到一个数，接着把这个数代入方程式中得到另一个数，用第二个数又得到第三个数，如此等等。其实，早在 20 世纪的最初几年，法国数学家亨利·普安卡雷（Henri Poincaré）就尝试着把这些连续的成对数值绘在图上，用这种方式理解一组复杂的微分方程式。普安卡雷在图中发现了一些值得关注的图式，却因不知道如何对这些图式做进一步的研究而放弃了深入研究的想法。而混沌理论就是以普安卡雷的这些图式为起点发展起来的。当你在绘制一张普安卡雷图形（Poincaré plots）时，会发现图纸上出现的那些点起初好像完全不成形状，表面上这些点以一种偶然的方式出现在随便什么地方，但承受着绘在图上的点数的不断增加，图式开始显现出来，有时是几组平行线，有时也可能是一组相互交叉的线，或许是很多个圆，或是和直线相交的圆。

---

[12] 此处有关混沌理论的描述，取自 Brian Davies，Exploring Chaos：Theory and Experiment（Reading，MA：Perseus Books，1999）一书。

混沌理论的拥护者认为，现实生活中那些看上去是纯随机的测量值，实际上是由某个确定性的方程组生成的，这些方程可以从普安卡雷图形的模式推演出来。例如，有些混沌理论的拥护者记录下了人类心脏动脉搏动的间隔时间，并绘成普安卡雷图形。他们声称在这些图上看到了一些形状，并且已经发现一些似乎能产生同类形状的确定性生成方程。

直到写这本书时为止，以这种方式应用的混沌理论仍存在着一个严重的缺陷。根据数据绘出的图形与用一组特定方程组生成的图形，这两者之间的拟合度如何，并未测量。他们只是要求读者观察两种相似的图形，并以此为依据证明给出的生成方程是正确的。统计分析上已经证明这种用肉眼检验的方式难免出错。因为，用肉眼判断类似的或几乎完全相同的两个图形，如果改用为此目的创建的统计分析工具仔细检验之后会发现，两者往往是大不相同的。

皮尔逊的假使优度检验

这是 K·皮尔逊在他的学术生涯早期就已经意识到的一个问题，K·皮尔逊最伟大的成就之一就是创造出第一个“拟合优度检验“(goodness of fit test)。通过观测值与预测值的比较，皮尔逊构造出一种能对拟合优度进行检验的统计量，并称之为“χ^2拟合优度检验”（chi square goodness of fit test)。之所以用希腊字母χ（读作“kai”），是因为这个检验统计量的分布属于一组偏斜分布，而他称这组偏斜分布为χ家族(chi family)。实际上，这个检验统计量很像χ的平方，因此命名为“χ^2”。在费歇尔看来，既然是一个统计量，就会服从一种概率分布。K·皮尔逊证明了无论用哪一种类型的数据，χ^2拟合优度检验都服从相同的分布。也就是说，他能列出这个统计量的概率分布表。每一个检验都能用到同样的那套表。χ^2拟合优度检验只有一个参数，费歇尔称之为“自由度”。费歇尔在在 1922 年的那篇论文里，首次批评了皮尔逊的研究，指出在比较两种比例时，皮尔逊得出的那个参数值是错误的。

但是，没有任何理由只因为皮尔逊理论上的一个很小的错误，就贬低他的这项伟大成就。皮尔逊的拟合优度检验是现代统计分析中一个重要组成部分的先驱，这个重要组成就是“假设检验”(hypothesis testing）或“显著性检验”(significance testing)，它允许分析人员提出用来模拟现实的两种（或多种）不一致的数学模型，然后利用数据来放弃其中的一个。假设检验应用得如此广泛，以至于很多科学家认为这是他们唯一能用的统计方法。在后面的章节中我们会发现，假设检验的应用甚至涉及到一些严肃的哲学问题。

检验女士是否真能品尝出茶的区别

假设我们要检验那位女士能否品尝出两杯茶的不同：是把牛奶倒进了茶水里，还是把茶水倒进牛奶里。我们给她两杯茶，告诉她一杯是茶水倒入牛奶里，另一杯是牛奶倒入茶水中。她尝了尝，正确区别开了这两杯茶。有可能她是凭猜测，猜对的机会是一半对一半。我们再给她同样的这样两杯茶，她又说对了。如果她仅仅靠猜测，那么连续两次都猜对的机会是四分之一。如果我们再给她两杯茶，假如她仍然能正确地分辨出来。若这人结果完全是猜出来的，此时猜对的机率则只有八分之一。我们继续两杯两杯地让她品尝更多杯茶，而她依然每次都能够正确地识别出来。某种意义上，我们就不得不相信她真的能品尝出其中的差别了。假定她说错了一次，假定说错的这一次就发生在第 24 组，而其他的全对，那么我们能否依然认为她真的有分辨不同奶茶的能力呢？假如她的错误是二十四分之四呢？或是二十四分之五呢？

假设检验（或者说显著性检验）是一种正规的统计方法，是在“待检验的假设为真”的假设前提下，用来计算以往观测到的结果发生的概率。当观测结果发生的概率很低时，我们

得出原假设不成立的结论。重要的一点是，假设检验提供了一种拒绝某个假设的工具。上述例子中，待检验的假设是：那位女士只是凭猜测。假设检验的目的不是让我们接受某个假设，即使与那个假设有关的概率非常高也不能接受。

在这个普遍被接受的概念发展的早期，“significant”（显著的）这个词只是用来指“概率低到足以拒绝的程度”，数据如果可以用来拒绝某个分布，则它就是显著的。在 19 世纪后期的英语里，这个词仅仅是指计算结果意味着或表明了什么意思。进入 20 世纪之后，英语“significant”这个词在原有含义的基础上又扩展了其他的解释意义，也指某些事情是非常重要的。在某个待检验的假设条件下，统计分析仍沿用“significant”这个词“显著的”含义来表示计算结果发生的概率很低，在这个层面上，“significant”这个词有一个精确的数学涵义。但令人遗憾的是，使用统计分析的人常把显著性检验统计量理解为某种更接近这个词的现代语意的东西。

费歇尔对 P 值的运用

现在运用的显著性检验方法，其中大部分都是费歇尔构造出来的。他把判定具有显著性的那个概率，称为“P 值”（P-value）。他对 P 值的涵义和有效性坚信不疑。在《研究工作者的统计方法》一书中，很多地方都专门介绍了怎么计算 P 值。正如我在开头的时候谈到的，这是一本专门给想要应用统计方法的非数学专业人士写的书。在这本书中，费歇尔并未解释这些检验是如何推导出来的，也从没有明确指出究竟多大的 P 值才算是显著的。他只是举出一些计算实例，并说明结果是否显著。在一个例子中，他给出一个小于 0.01 的 P 值，并且说明“一百个值当中，只有一个值会偶然超过（计算出来的检验统计量），因此，很显然，计算结果之间的差异具有显著性。”

1929 年，费歇尔在《心灵研究学会刊》（Proceedings of the Society for Psychical Research）上发表的一篇论文中，几乎等于定义了一个在任何情况下都将是显著的特殊的 P 值。“心灵研究”（psychical research）提到试图用科学的方法来证明“超视力”的存在。心理学的研究人员大量运用了统计学的显著性检验来证明，在受实验者完全随意猜测这种假设条件下，其结果是不可能的。费歇尔在他这篇论文中，先是谴责某些作者完全错误地使用了显著性检验，接着他申明说：

> 运用生物学的方法对生物界进行观察的时候，统计学的显著性检验是必不可少的。其作用就在于防止我们被一些非主要的偶发事件所欺骗。并不是因为我们希望去研究或试图去查明这些偶发事件，而是因为它们与许多我们无法控制的其他境况联系在一起。一个观测的结果，倘若在我们正在寻找的真正原因根本不存在的情况下，几乎从未发生过，可以判断这个观测具有显著性。如果偶然发生的机率低于二十分之一，通常的做法是判断其结果具有显著性。对实际调查者来说，显著性水平的选择是任意的，但便于应用。不过，它并不意味着可以让自己每 20 次实验中就被骗一次。显著性检验只是告诉他什么是应该忽略掉的，也就是说应该把所有那些无法得到显著性结果的实验忽略掉。当他知道如何设计一个实验，而这个实验几乎一定能给出一个显著性的结果时，他也只能说明，这仅是一种实验上可以验证的现象。所以，对那些孤立的具有显著性的结果，他不知道如何才能让它们再现出来，只能留待以后再做进一步的调查研究了。

注意“……知道如何设计一个实验，而这个实验几乎一定能给出一个显著性的结果……”这句话，正是费歇尔使用显著性检验的核心之所在。对费歇尔而言，显著性检验只有在连续实验的相互联系中才有意义，所有这些实验的目的在于解释特定处理的作用。读过费歇尔的应用性论文之后，你会在他的引导下相信，使用显著性检验是为了得出三种可能的结论之一：如果 P 值很小（通常小于 0.01），他断言某种结果已经显现出来；若 P 值很大（通常大于 0.2），

他宣称即便真的存在一个结果，也会因为该结果发生的可能性太小，所以不可能有任何显示出这个结果的大规模的实验；如果 P 值介于前两者之间，他讨论了应该如何设计下一个实验，才能得到一个更好的结果。除了上述情况，费歇尔从来没有明确说明科学家应该怎么解释 P 值。对费歇尔而言，看上去是如此显而易见的事，对读者来说可能并不清楚。

我们将在第 18 章回过头来重新审视费歇尔对显著性检验的态度。费歇尔始终坚持，从来都没有显示过吸烟有害健康，这也正是他的一个较大错误的核心之所在。费歇尔对有关吸烟和健康的证据做了犀利的分析，我们暂且把它放下，以后再谈。现在把话题转到 1928 年，看看当时 35 岁的耶日·奈曼。

J·奈曼的数学教育

当第一次世界大战在东欧爆发，奈曼的祖国陷于战火之中的时候，他还是一个在数学系读书的非常有发展前途的大学生。他被近搬到俄国，就读于卡尔可夫大学（University of Kharkov）——一个远离数学活动的视野偏狭的地方。学校缺少具有当代最新数学知识的合格老师，而且由于受到战争的影响，他是在学期中途才入学的，因此，在卡尔可夫，他只学到一些最基础的数学知识。奈曼只能寄希望于那些能得到的数学期刊，从中查找论文文献。可想而知，奈曼受到的正规的数学教育只相当于 19 世纪学生学到的内容，20 世纪的数学知识则是他通过自学掌握的。

对奈曼来说，可利用的数学期刊仅限于卡尔可夫大学的图书馆和后来在当地的波兰学校图书馆里能找到的。偶然的机会，他发现了亨利·勒贝格（Henri Lebesgue 1857-1941）的一套论文集。20 世纪的最初几年，勒贝格提出许多现代数学分析的基本思想，但是他的论文晦涩难懂。后来的数学家把勒贝格积分、勒贝格收敛定理以及这个伟大数学家的其他一些创见简化并整理成更容易理解的形式。现在已经没有人再去读勒贝格的原谅了，学生们都是通过阅读这些新版的文章来学习勒贝格的思想。

所谓的“没有人”当然是除了奈曼之外的，当时他只有勒贝格的原文可以读，他苦读这些原文，从中感受到了这些全新的（对他而言）伟大创见所蕴含的辉煌。此后的许多年，奈曼一直非常景仰勒贝格，20 世纪 30 年代末在法国的一次数学研讨会上，终于得以与勒贝格见面。据奈曼所说，勒贝格表现得态度生硬、粗鲁无礼。当奈曼热情洋溢地表达对他的仰慕时，他阴郁冷淡地回应了一句，就转身离开了正在喜出望外地等待与他交谈的奈曼。

这种冷淡让奈曼深受伤害，并且，奈曼可能把这次经历当作了反面教训，他对青年学生一直都格外的亲切有礼，仔细地倾听他们的谈话，并对他们的热情给予鼓励和回应。奈曼正是这样的一个人。所有认识他的人都对他的亲切和蔼、富于同情心的为人记忆犹新。他与人为善、体贴入微、待人真实宽厚。当我见到他的时候，他已经 80 多岁了，一个身材瘦小、举止高贵、衣着讲究、蓄着整洁白胡须的老人。他在听别人讲话和别人深入交谈的时候，蓝眼睛神采奕奕地闪烁着，对每个人都同样地全神贯注，无论对方是谁。

在他的职业生涯之初，奈曼好不容易才找到工作，成为华沙大学（the University of Warsaw）的一个年轻的教师。当时，刚刚独立的波兰因资金短缺，没钱资助学术研究，也很少有给数学家的职位。1928 年，他在伦敦的生物统计实验室呆了一个暑假，并认识了 E·皮尔逊和他的太太艾琳（Eileen）以及他们的两个女儿。E·皮尔逊是 K·皮尔逊的儿子，但是父子两人在个性上的天壤之别可谓绝无仅有：K·皮尔逊精力充沛，有支配控制他人的欲望；E·皮尔逊却腼腆谦虚。K·皮尔逊喜欢追逐新观念，常在数学概念还相当模糊，甚至还存在某些错误的时候，就忙着发表论文；E·皮尔逊则极其小心谨慎，甚至为每一步计算的细枝末节担忧。

E·皮尔逊与奈曼的深厚友谊长存在两人 1928-1933 年间的通信中。这些信件展示了他

们对社会科学卓越的洞察力，以及两颗富于独创精神的心灵是如何提出各自的想法，或批评对方的想法，并共同解决难题的。E·皮尔逊踌躇地指出奈曼的提议或许不可行，这时他表现出谦逊的一面；奈曼巧妙地剖析复杂的问题，并抓住每个难题的重要本质，这时展现出他的独创力。有人如果想知道数学研究为什么是需要经常进行合作的事业的话，我建议他看看奈曼与 E·皮尔逊的通信。

E·皮尔逊对奈曼提出的第一个问题是什么呢？回想 K·皮尔逊的 x^2 拟合优度检验，他创立这种方法来检验观测数据是否与理论分布相符。但事实上根本不存在像 x^2 拟合优度检验的这种东西。分析人员有无数种方法可用来对给定的一组数据进行检验，似乎没有任何准则能够判定如何在这么多的选择中挑选出“最好的”。每次用到检验的时候，分析人员必须做出一个相当随意的选择。对此，E·皮尔逊问了奈曼以下的问题：

如果我用了 x^2 拟合优度来检验一组服从正态分布的数据，但我没能得到一个显著的 P 值，那么我怎么知道这组数据确实服从正态分布呢？也就是说，我怎么知道至今尚未发现的另一种 x^2 检验或者另一种拟合优度检验不会已经产生了一个显著的 P 值，而允许我在拟合数据的时候拒绝这个正态分布呢？

奈曼的数学风格

奈曼把这个问题带回华沙，并由此而开始了两人之间的书信往来。奈曼与小皮尔逊都对费歇尔建立在似然函数基础上的估计概念印象深刻。通过检查与拟合优度检验联系在一起的似然函数，他们开始了调查研究。两人联名发表的第一篇论文介绍的就是那些研究的结果。这是他们撰写的三篇顶尖论文当中最难的一篇，它几乎彻底变革了关于显著性检验的全部思想。当他们继续探索这些问题时，奈曼极度清晰的洞察力使问题在蒸馏中不断提纯，精炼出最基本的元素，使他们的研究成果变得更为清晰，也更容易理解。

虽然读者对此可能不太相信，但在数学研究领域，一个人写文章的风格确实发挥着很重要的作用。有些数学文献的作者似乎写不出让人容易理解的文章；有些人则似乎以写成一行又一行的数学符号与注释为乐事，一篇论文中充斥着无比繁琐的细节，以至于把总的思考都迷失在了微不足道的细节中。与之相反，有些作者却总是有能力用非常简单而有说服力的方式表达复杂的思想，数学的发展在他们的表达中显得如此的鲜明而平实。只有在回顾已经学到些什么时，读者才会确实认识到结果的伟大力量。奈曼就是这样的作者，读他的论文是件令人愉快的事，数学观点自然地展开，使用的符号简单得令人无法相信，结论的显现竟如此的自然，以至于让人感到难以理解，不禁要问，为什么很久以来居然没有人发现这项结论？

我在辉瑞的研究中心工作了 27 年，该中心每年都赞助康涅狄格大学举办一次学术年会。该校的统计系通常会邀请一位生物研究方面的重要人物来一天，与学生们见面聊聊，随后，会在下午的晚些时候发表演讲。由于我曾经参与负责一年一度的研讨会的资金事宜，因此有幸会见统计学界的一些大人物，奈曼就是应邀者之一。在一次研讨会前，奈曼想让他的演讲以一种特殊的方式进行，他先介绍一篇论文，随后组织一个专题组来评判他的论文。由于是大名鼎鼎的奈曼，研讨会的组织者联系了美国新英格兰地区著名的资深统计学家组成了这个专题讨论组。在研讨会开幕前的最后一记得，有位专题组成员无法出席，于是会议安排我代替他。

奈曼事先已经把他打算演讲的论文印发给了我们。那真是篇激动人心的论文！论文中奈曼利用他 1939 年完成的研究成果，去解决一个天文学上的难题。我知道 1939 年的那篇论文。几年前，当我还是个研究生的时候就看到了它，并留下了深刻的印象。论文中阐释了奈曼已经发现的一类新的分布，他称之为“散播分布”（contagious distribution）。论文中所提到的问题，开始是试着模拟土壤里昆虫幼虫的分布情形：即将排卵的母昆虫带着满肚子的卵

在田野里四处飞，然后随机选取一个地点排卵，一旦排完卵，幼虫孵化出来，就从那个地点钻出地面。现在，从田野里取一个土壤样本，那么，在这个样本里发现的幼虫数量的概率分布是什么？

散播分布描述了这种情形。奈曼 1939 年的论文，运用一系列看似简单的方程，导出散播分布。推导的过程看上去明显而自然。显然，看完论文之后，读者会觉得除了奈曼的做法之外，再没有更好的推导方法了。但这只是在读了奈曼的文章后才清楚的。自从 1939 年那篇论文发表之后，人们发现奈曼的散播分布适用于相当多的领域，如医学研究、冶金术、气象学、毒物学，以及解决宇宙中星系的分布问题（就像奈曼在辉瑞的那个研讨会介绍论文所描述的）。

演讲结束，奈曼坐下来听专题小组的讨论。讨论组的其他成员都是著名的统计学家。由于太忙，不能提前阅读他的论文，他们把辉瑞的研讨会作为对奈曼荣誉的肯定。他们的“讨论”包括对奈曼的学术生涯和以往建树的评论。我作为最后一记得的替补者加入到这个专题组中，并且被告之不能提及我先前和奈曼相处的经历（其实我根本没有这种经历）。因此，我就应他的本意，直接评论奈曼那天演讲的东西。我提到在几年前是如何发现了 1939 年的那篇论文，以及为了准备参加座谈会，重读了论文。我尽一切所能描述论文的内容，谈到奈曼创立的分布参数其意义的巧妙方式时，我显出极大的兴趣。

奈曼对我的评论显得非常高兴。之后，我们俩热烈地讨论了散播分布以及它的用法。几周以后，我收到寄来的一个大包裹，是一本加州大学出版社（The University of California Press）出版的《J·奈曼早期统计论文选》（A Selection of Early Statistical Papers of J. Neyman），在书的内封有一行题词：“致大卫·萨乐斯伯格（David Salsburg）博士，衷心感谢他在 1974 年 4 月 30 日对我演讲的有趣讲评。J·奈曼。”

我把这本书视为珍宝，一是由于奈曼的题字，二是因为书中那一系列精美绝伦、文笔极佳的论文。从那时起，我有机会与奈曼的很多学生和同事交谈，得知这个我在 1974 年碰到的、友善的、风趣的、有感召力的人，也是他们深知并崇敬的人。

第 11 章 假设检验

在他们一开始合作的时候，E·皮尔逊就问耶日·奈曼，在检验一组数据是否为正态分布时，如果没能得到一个显著性的 P 值，那么怎样才能看这组数据是正态分布的呢？他们的合作从这个问题开始，然而，E·皮尔逊最初的这个问题，却打开了一扇通往更广阔领域的大门。在显著性检验中，如果得到的是一个不显著的结果，那么它的涵义是什么呢？如果我们找不到拒绝一个假设的证据，我们能做结论说这个假设为真吗？

费歇尔其实已经间接地回答了这个问题。费歇尔把比较大的 P 值（代表没有找到显著性证据）解释为：根据该组数据不能做出充分的判断。依据费歇尔的解释，我们绝对不会得出这样的推理，即没有找到显著性的证据，就意味着待检验的假设为真。这里引用费歇尔的原话：

相信一个假设已经被证明是真的，仅仅是由于该假设与已知的事实没有发生相互矛盾，这种逻辑上的误解，在统计推断上是缺乏坚实根基的，在其它类型的科学推理中也是如此。当显著性检验被准确使用时，只要显著性检验与数据相矛盾，这个显著性检验就能够拒绝或否定这些假设，但该显著性检验永远不能确认这些假设一定是真的，……如果显著性检验真的被人们理解到这种程度，那么就说明显著性检验的道理已被人们认识清楚了……

在这之前，K·皮尔逊常常利用他的卡方拟合优度检验来“证明”某些数据符合某些特定的分布。在费歇尔把更精确的方法引入到数理统计之后，K·皮尔逊的方法就不再为人接受了。但问题仍然存在。为了知道应该估计哪些参数，为了确定这些参数与所研究的科学问题之间有何关系，我们必须假设该数据符合某一特定的分布。统计学家们常常会利用显著性检验来证明数据符合何种分布。

在他们的通信往来中，E·皮尔逊与奈曼经常探讨一些由显著性检验中浮现出来的悖论，不假思索地使用一项显著性检验，可能会把一个显然为真的假设拒绝掉。但费歇尔从未陷入这种尴尬，因为对他来说，显著性检验怎样被误用他是非常清楚的。奈曼问：用什么标准来判断一项显著性检验的应用是正确的还是不正确的呢？逐渐地，随着 E·皮尔逊与奈曼的书信往来，加上奈曼在暑期到英国的几次访问以及 E·皮尔逊的几次波兰之旅，假设检验的基本思想已经浮出水面[13]。

现在，在所有基础统计学的教科书中，都可以发现一个简化的奈曼－皮尔逊假设检验理论公式。该公式结构简单，我发现大部分的大学一年级学生很容易看懂，因为已经被编纂整理过，所以这个公式很精确，也很有说服力。假设检验理论必须这样来写，当然这也是教科书所需要的写法，也只能这样来写。这种直接表述假设检验的方法已经被一些政府和社会机构所接受，如美国食品及药品管理局、美国环保署，许多医学院在给将来做医学研究的人授课时，采用的也是这一套方法。此外，这种方法也逐渐地被应用到了司法界，当法院处理某些需要鉴别的歧视性案子时，就经常会用到这种方法。

当由奈曼和 E·皮尔逊创建起来的这种理论以奈曼的这种直接而简化的方式来讲授时，由于集中于公式中有错误的一面，从而曲解了他的发现。奈曼的主要发现是，除非至少有两个可能的假设，否则显著性检验根本就没有意义。也就是说，你不可能检验一组数据是否服从正态分布，除非你认为该组数据也可能会被其它的一些分布或分布集来拟合。这些备择假

[13] 整个这一章，我把核心的数学观点归功于奈曼，因为奈曼不但很仔细地导出了相关的数学公式，也负责这些公式表述的最后润色。但是在 E·皮尔逊与奈曼见面之前 6 个月，E·皮尔逊就曾与威廉·西利·戈塞特通过信，这表明 E·皮尔逊已经思考过备择假设以及不同类型误差的问题了，而戈塞特可能是最先提出这个想法的人。尽管 E·皮尔逊是首先提出该问题的人，但他也承认是奈曼为他的“粗略想法”提供了坚实的数学基础。

设的选择，决定了显著性检验的执行方式。当一个备择假设为真时，该备择假设被接受的概率奈曼称之为该检验的效力（power）。在数学里，要清晰阐述一种思想，通常要给某一特定的概念赋予清楚明确的定义。为了区别被用来计算费歇尔 P 值的假设与其它可能的一个或多个假设，奈曼和 E·皮尔逊把被检验的假设称为“零假设”（null hypothesis），称其它可能的假设为“备择假设”（alternative hypothesis）。在他们的理论公式中，计算 P 值是为了检验零假设，而检验的效力则是指在备择假设为真的条件下 P 值的表现效果。

奈曼由此得出两个结论。第一个结论是，检验的效力是用来测量一个检验方法好坏的指标，两种检验方法中效力较强的方法就是较好的方法；第二个结论是，备择假设不能太多。统计分析师不能这样来表述，某一组数据来自于一个正态分布（零假设），或者它来自于任何其它可能的分布。这种备择假设集涵盖的范围太广了，没有哪种检验方法会有那么强的效力能处理所有可能的备择假设。

在 1956 年，芝加哥大学的 L·J·萨维奇与拉杰·拉克·巴哈杜尔（Raj Raghu Bahadur）证明，对于一个零假设未通过的情形，并不一定要求有很多的备择假设。他们构建了一个相对较小的备择假设集，除此之外的所有检验的效力均为零。在 20 世纪 50 年代，奈曼就发展出了有限制的假设检验的想法，其中的备择假设集被定义得非常狭窄。他证明得出了这样的结论：这种检验方法比那些处理较多备择假设的检验方法效力更强。

在很多情况下，假设检验的目的是用来推翻零假设的，而这个零假设就好比我们所要攻击的稻草人。举例来说，当我们比较两种药的临床效果时，待检验的零假设是两种药的效果一样。但是，如果真是如此，研究工作就永远不必进行了。所以，“两种处理的效果相同”这一零假设，就是我们所要攻击的稻草人，应该被我们研究的结果来推翻。因此，根据奈曼的思想，该项研究的设计，该项研究的设计必须使最终数据有最大的检验效力，这样才能推倒这个稻草人，即表明这两种药的效果有多大的不同。

什么是概率？

遗憾的是，为了对具有内部一致性的假设检验设计出一种数学方法，奈曼必须处理一个已被费歇尔扫到地毯下的问题。这是一直困扰假设检验的一个问题，尽管奈曼的纯数学解非常简洁巧妙。这也是统计方法应用到一般的科学领域中通常会碰到的问题。从更一般的意义讲，这个问题可以这样来概括：在现实生活中，概率的意义是什么？

统计学的数学公式可用来计算概率。而这些计算出来的概率可使我们应用统计方法解决科学中的问题。就所用到的数学而言，概率的定义很明确。但这种抽象的概念怎样和现实相联系呢？当科学家试图决定什么为真、什么不为真时，他该如何解释统计分析的概率陈述呢？在本书的最后一章，我将讨论这个一般性的问题，并分析长久以来设法解答这些问题所做的努力。但现在，我们将分析促使奈曼找到他的答案的特殊情况。

前面我们谈过，费歇尔利用显著性检验产生了一个他称为 P 值的数字。这是一个计算出来的概率，是在零假设为真假定下，与观测数据有关联的一个概率。例如，假定我们要检验一种新药，对做过乳房切除手术的妇女来说，这种药可以防止乳腺癌的复发。我们把这种药的效果与一种安慰剂作比较。此时的零假设（那个稻草人）就是，该新药不比安慰剂好。现在，假定 5 年之后，用安慰剂的妇女有一半乳腺癌复发，但用新药的完全没有复发，这样能证明新药“有效”吗？答案当然得看这个 50%代表多少病人。

如果在这项研究中，两组各仅有 4 名病人，也就是总共有 8 名病人，而其中 2 人在 5 年后复发。假定我们任选一个 8 人团体，把其中两人做上标记，接着把人随机分成两组，每组 4 人，那么做标记的两个被分在同一组的概率大约是 0.30. 因此，如果每组只有 4 名妇女，“所有复发的妇女都落在安慰剂组”是不显著的。如果该项研究中每一组包含 500 名妇女，

且乳腺癌复发的所有 250 名妇女都落在安慰剂姐，这是极度不可能的，除非新药真的有效。如果新药并不比安慰剂有效，这 250 名妇女都落在同一组的概率就是 P 值，计算出来的结果将小于 0.0001.

P 值是一个概率，它就是这样被计算出来的。既然 P 值被用来表明一个假设（P 值就是在该假设下计算出来的）为假的概率，那它的实际意义又是什么呢？答案是，P 值是在极可能为假的条件下，与观测值相关联的一个理论概率。P 值与现实没什么联系，它是一种对似是而非问题的间接测量。它不是我们错误理解的新药有效的概率，它也不是出现任何一种类型误差的概率。但是，为了决定哪一种检验方法比别的检验方法更好，奈曼必须想出一种办法把假设检验放进一个架构里，使得与根据检验所做出的决策相联系的概率能够计算出来的。因此，他需要将假设检验的 P 值与现实生活联系起来。

概率的频数定义

1872 年，英国哲学家约翰·维恩（John Venn）提出了一个数学概率的公式。这个公式使得概率在现实生活中有了含义。他把一个重要的概率定理转了一个方向，这个定理就是大数定律（law of large numbers）。大数定律指出，如果某事件有给定的概率（比如掷一个骰子，得到六点这一事件的概率是六分之一），而且如果我们重复地进行相同的试验时，该事件发生的次数的比率就会越来越接近这个概率值。

维恩指出，与一个给定事件相联系的概率，是该事件从长期来看所发生的次数的比率。按照维恩的意见，概率的数学理论并没有隐含大数定律，反而是大数定律隐含了概率的思想。这就是以频数为基础对概率的定义。1921 年，约翰·梅纳德·凯恩斯（John Maynard Keynes）[14]推翻了这种定义方式，认为它不是一种有用的或有意义的解释，并指出这种定义具有根本性的矛盾，因而无法在许多要求计算概率的情况不应用概率的频数定义。

在用正规的数学方法来构造假设检验时，奈曼又重新回到了维恩的概率的频数定义上。奈曼利用这个定义来证明他在假设检验中对 P 值解释的合理性。在奈曼－皮尔逊的公式中，科学家设定一个固定的值，比如 0.05，之后，当显著性检验的 P 值小于或等于 0.05 时，就拒绝零假设。按照这种理解，从长期来看，该科学家会正好有 5%的机会拒绝一个正确的零假设。假设检验当前就是这样来讲授的，奈曼所采用的频数方法被得到强调。我们太容易把奈曼－皮尔逊的假设检验公式看作是概率的频数方法的内容，因而太容易忽略奈曼所提的观点中更重要的见解，即为了检验零假设这个“稻草人”，必须要有一组定义明确的备择假设。

费歇尔误解了奈曼的见解。他把注意力集中到了显著性水平的定义上，但却忽略了检验效力和需要定义一组备择假设这些重要的思想。在批评奈曼时费歇尔写到：

> 奈曼认为他自己修正并改善了我早期所做的关于显著性检验的工作，结果“改进了自然知识”，不过实际上他只是用技术性与商业性的形式，也就是大家所熟知的接收程序，重新解释了这些检验方法罢了。现在，在当代世界里，这种接收程序变得十分重要。例如，当英国海军总部接到某工程公司的大批材料时，我认为要安排很仔细的检查与检验，以降低残次品被接收的频率，……不过在我看来，这种管理运作与透过物理或生物

[14] 从某方面来说，误称定律也发生在 J·M·凯恩斯身上。很多人都认为他是经济学家，凯恩斯经济学派的创始人。凯恩斯经济学派认为，政府可以通过货币政策的调控来影响一个国家的经济发展过程。不过凯恩斯拿的却是哲学的博士学位，而他在 1921 年出版的博士论文，题目竟然是《关于概率的讨论》（A Treatise on Probability），是探讨数理统计应用背后的哲学基础发展的重要代表作。在本书往后的章节里，我们还将引用到凯恩斯说的话。不过，我们是把凯恩斯作为一个概率学家来引用他的话，而不是把他作为一个经济学家来引用的。

实验的科学发现工作相比，它们之间的逻辑上有很大的差别，所以拿这两者做类比是没有多大帮助的，而把它们当成是同一回事，更是一种决定性的误导。

尽管存在对奈曼基本观点的这些扭曲，假设检验还是成为科学研究中应用得最多的统计工具。奈曼提出的精巧数学构思，在科学的很多领域中都占有一席之地，变成了一种固定的观念。大部分的科学期刊都要求论文的作者在做数据分析时要采用假设检验方法，甚至连科学期刊之外的领域也开始这么做。美国、加拿大与欧洲的药物管理机构，纷纷把假设检验方法的使用列为对药品检查的强制性要求，就连法庭允许原告用这种方法证明自己受到就业歧视。假设检验已经渗透到统计学的所有分支学科中。

奈曼－皮尔逊的理论攀升到统计学的巅峰地位，一路上也不是没有挑战的。费歇尔从一开始就攻击它，而且在他有生之年一直在攻击这个理论。1955 年，费歇尔在《皇家统计学会期刊》上发表一篇文章，题目是“统计方法与科学归纳”，而在他的最后一本书《统计方法与科学推论》（Statistical Methods and Scientific Inference）里，更进一步详述了他的看法。在 20 世纪 60 年代晚期，不久之后就出任《生物统计》期刊主编的大卫·考克斯（David Cox），发表了一篇分析清晰的文章，分析了假设检验在科学中的实际用途，同时也证明了奈曼的关于频数的解释不符合实际状况。在 20 世纪 80 年代，W·爱德华兹·戴明（W. Edwards Deming）攻击了假设检验的整个思想，认为假设检验的整个思想都是荒谬的（第 24 章还会再提到戴明对统计学的影响）。年复一年，在统计学文献中一直有相关文章发表，指出在教科书中已成定格的奈曼－皮尔逊理论中发现了新的毛病。

不过，在奈曼－皮尔逊假设检验理论的神圣化过程中，奈曼本人并没有参与。早在 1935 年，他在《法国数学学会会刊》《bulletin de la Société Mathématique de France》上就用法文发表过一篇文章，对是否能找到最佳的假设检验方法提出严厉的质疑。在他后来的文章里，奈曼很少直接使用假设检验方法，他的统计方法通常是由理论原则导出概率分布，然后再由数据来估计参数。

其他一些人则捡取藏在奈曼－皮尔逊理论背后的观点来进一步发展。在第二次世界大战期间，亚伯拉罕·沃尔德扩展了奈曼利用维恩关于频数的定义，发展成了一个叫统计决策理论（statistical decision theory）的领域。埃里希·莱曼（Erich Lehmann）给出了用来判断一个好的假设检验可供选择的标准，后来在 1959 年，他还写了一本有关假设检验问题的权威性的教科书，这本书至今仍然是该领域对奈曼－皮尔逊假设检验理论描述得最完整的一部著作。

就在希特勒入侵波兰，将邪恶之幕笼罩欧洲大陆之前，奈曼就到了美国，并在加州大学的伯克利分校开始创建统计系。在那里他一直工作到 1981 年去世，这期间，他把该系创建成全世界最重要的学术性统计学系之一。他把一些统计学界赫赫有名的人物引入该系，同时也提拔了一些默默无闻的人，这些人正致力取得卓越的成就。例如，大卫·布莱克韦尔（David Blackwell）原来只是只身孤单地在霍华德大学（Howard University）工作，没有数理统计同行与他来往。由于他的种族原因，他一直没能在“白人”学校谋得一职，尽管他很有潜能。奈曼把他请到了伯克利。此外，奈曼还招了一位出身法国农民家庭的研究生吕西安·勒卡姆（Lucien Lecam），他后来成为世界领先的概率学家。

奈曼总是非常和善地对待他的学生和同事。他们常常津津乐道的是系里每天下午茶歇的欢乐时光，这是由奈曼主持的他与职员亲近接触的一个重要场合。他总是亲切地鼓励学生和同事谈谈自己最新的研究成果，同时很和蔼地提出他自己的思路和见解，给出评论，加入大家的讨论。他常常在下午茶歇即将结束时举起茶杯说“为尊敬的女士们！”他特别关照女士，鼓励她们在学术生涯上不断进步。在他的女弟子当中，伊丽莎白·斯科特（Elizabeth Scott）博士是较为杰出的，她与奈曼一起做研究，共同发表论文，范围从天文学到致癌物研究，甚至动物学。还有伊夫琳·菲克斯（Evelyn Fix）博士，她在流行病学的研究上有很重要的贡

献。

直到费歇尔于 1962 年去世，奈曼一直受到这位天才的尖刻批评。奈曼每做一件事都会遭到费歇尔的批评。如果奈曼成功地证明出了费歇尔某项非常难解的叙述，费歇尔就说奈曼误解了他写的东西；要是奈曼扩充了费歇尔的某个观点，费歇尔就批评奈曼说他把好端端的理论用错了地方。对比，不论是付诸笔端，还是在私人场合，奈曼从不回应（如果我们相信奈曼同事的说法）。

在奈曼去世前的一次访谈中，奈曼说了一件发生在 20 世纪 50 年代的往事。当时他准备在一次国际研讨会上公开发开一篇用法语写的论文。当他步上讲台时，意识到费歇尔也坐在听众席上。在演讲论文时，他知道一场激辩难免，于是开始武装自己，他预计费歇尔会抓住论文里某个无关紧要的小地方，将论文和他本人攻击得体无完肤。奈曼讲完之后，等待听众提问，结果只有几个问题。费歇尔相当平和，一言未发。后来奈曼才知道，费歇尔不会讲法语。

第 12 章 置信诡计

当 20 世纪 80 年代出现了艾滋病（AIDS）这种传染病时，有若干问题需要回答。一旦传染源 HIV（human immunodeficiency virus，即人体免疫缺损病毒）确定了，卫生官员需要知道有多少人受到感染，以便安排需要的资源来应付这种传染病。幸运的是，在此之前的 20 至 30 年所开发出来的流行病学[15]数学模型，在这里可派上用场。

从传染病的现代科学观点来看，某些个体病人接触到传染源，其中有些人会被传染，而在经过一段所谓的“潜伏期”之后，那些被传染的人会显现该疾病的症状。一旦被传染，这个人就会成为其他还没有被传染人的潜在传染源。我们没有办法预测谁会与传染源接触，谁会被传染，或谁会传染他人。我们所能做的，只是处理相关的概率分布，并估计这些分布的参数。

参数之一是平均潜伏期，也就是从被传染到症状产生的平均时间。就艾滋病这种传染病来说，平均潜伏期对卫生官员是特别重要的参数。他们没有办法知道究竟有多少人被传染，又有多少人最终会得上这种疾病，但如果能知道平均潜伏期，他们就能根据已经患有这种疾病的人数，估计出受感染的人数。不仅如此，由于艾滋病传染模式的不寻常特征，卫生官员拥有一组患者，并知道这组患者感染的时间和他们的发病时间。有一个小的血友病患者群体由于使用了被污染的血液制剂而感染上 HIV，他们提供的数据可以用来估计平均潜伏期这一参数。

这个估计值的准确性如何？流行病学家可以说，他们使用的是费歇尔意义上的最佳估计量。因为他们所得的估计值是一致的，又是最有效的。他们甚至还可以修正可能的偏差，并宣称他们的估计值是无偏的。但是，如果我们在前面章节里指出的，我们没有办法知道某一个具体的估计是否正确。

如果我们不能够说某个估计值是绝对准确的，那么我们还有没有办法可以说这个估计值与参数的真值之间有多接近呢？这个问题的答案在于使用区间估计（interval estimate）。点估计（point estimate）是一个单一的数字。例如，我们可能利用从血友病研究那里得到的数据，估计出平均潜伏期是 5.7 年。而一个区间估计会这样表述：平均潜伏期在 3.7 年至 12.4 年之间。在很多情况下，有区间估计的数字就够了，因为所需要的公共政策对区间估计的两端边界值来说是一样的。但有些时候，区间估计值显得太宽了，对最小的边界值和最大的边界值需要制定不同的公共政策。根据一个很宽的区间估计值所能得出的结论是，利用已有的信息不足以做出充分的决策，应寻求更多的信息，可以通过扩大调查的范围或进行一系列其它的实验来得到。

举例来说，如果艾滋病的平均潜伏期长达 12.4 年，则感艾滋病毒的人当中约有五分之一的人在感染之后要存活 20 年以上；如果平均潜伏期是 3.7 年，那么几乎每一个被感染的人在 20 年内都会发病。这两个结果相差太大。没有任何一种最佳的公共政策可以兼顾，因此需要更多的信息。

在 20 世纪 80 年代末期，美国国家科学院（National Academy of Science）如今国内一批顶尖的科学家组成一个委员会，讨论臭氧层破洞的问题。臭氧层可保护人类不受紫外线辐射的伤害，但由于人类使用的喷雾剂中含氟氯碳化物，可能破坏外层空间的臭氧层。这个委员会（主席为约翰·图基（John Tukey），是本书第 22 章讨论的主角）不做是或否的二分法回答，而是决定以概率分布的形式建立氟氯碳化物对臭氧层的影响模型。于是，他们计算出了臭氧层每年平均变化的区间估计值。虽然使用的数据量不是很多，但他们发现，该估计

[15] 流行病学（epidemiology）是与统计学关系非常密切的一门学科领域，其中统计模型被用来检验人类的健康模式。流行病学最简单的形式，是提供生命统计数据表以及相关统计分布参数的简单估计值。流行病学较复杂的形式，则是利用高级的统计理论，检验并预测流行疾病的发展进程。

区间的下边界值暗示，每年臭氧层将以一个较大的幅度减少，而这将使人类的生命在 50 年内受到严重的威胁。

区间估计现在已经普及到几乎所有的统计分析领域。当一项民意调查指出 44%的一般民众认为总统干得不错时，通常会加上一个附注，说明这个数字“具有正负 3 个百分点的误差”。上述民意调查结果的意思是，44%被调查的民众认为总统干得不错。由于这是个随机的调查，所求的参数是全国所有的民众中认为总统干得不错的人数的百分比。由于样本的容量较小，因此一个合理的猜测是，总体的参数值应落在 41%（44%－3%）与 47%（44%+3%）之间。

怎样计算区间估计值？怎样解释一个敬意估计值的涵义？我们能对一个区间估计值做出相应的概率表述吗？我们有多大的把握确信总体参数的真值会落在所估计的区间里？

奈曼的解

1934 年，耶日·奈曼在皇家统计学会做了一个演讲，题目是“论代表性方法的两个不同方面”（On the Two Different Aspects of the Pepersentative Method）。他的论文是关于抽样调查分析的。正如奈曼作品的一贯风格，这篇文章非常优美，导出了形式简单具直观易懂的数学表达式（当然是经过奈曼的推导之后才会如此）。但全文最重要的部分却在附录里，奈曼在这个附录中提出了一个很直接的方法，用来创建区间估计，并确定所得的区间估计值有多准确。奈曼称这个新的方法为“置信区间”（confidence intervals），而把置信区间的两端称为“置信界限”（confidence bounds）。

G·M·鲍利（G. M. Bowley）教授是大会的主席，起身致谢辞。他先用几段话讨论了论文的主要部分。接着就说到了附录：

> 我不太确定是否应该要求给出一个说明，或者直接提出质疑。论文的字里行间暗示，论文很难读懂，而我可能是被这个暗示误导的人之一（在这段话之后，他举出一个例子，表明他完全理解了奈曼提出的方法）。我只能说，从我一看到这篇论文开始，我就很认真地读它，而且昨天我还很仔细地读了奈曼博士对这篇论文的补充资料。我指的是奈曼博士的置信界限。我不太有把握地说，这里的“置信”是不是一个“置信诡计”。

鲍利接着举了一个例子说明奈曼的置信区间，然后继续说道：

> 这个方法真的会将我们引向深入吗？我们会比艾萨克·托德亨特（Isaac Todhunter，一位 19 世纪末的概率学家）知道的更多吗？它会让我们超越 K·皮尔逊和埃奇沃思（Edgeworth，数理统计发展早期的先驱之一）吗？它真的会引领我们到我们所需要的地方去吗？就是说我们所从中抽取样本的总体其比重会正好落在这些界限内吗？我看并不见得，……我不知道我是否已把我的想法表达清楚了，……自从我看到这个方法，我就觉得它是个难题。其理论陈述没有说服力，除非有人能说服我，否则我还是怀疑它的有效性。

鲍利对置信区间这个新方法的疑惑，是自从置信界限的概念被提出来以后大家对它的普遍迷惑之一。显然，奈曼在推导其结果过程中所用的四行优美的微积分式子，在抽象的概率数学理论上是正确的。它也确实能算出一个概率值。但这个概率值究竟代表什么则并不清楚。数据是观测得来的，参数是固定的值（尽管是未知的），因此参数取某个特定值的概率只有两个结果，或者是 100%（如果它就是那个值），或者是 0（如果它根本不是那个值）。然而，一个 95%的置信区间涉及的是 95%的概率。这个概率指的是什么？奈曼在此绕过了这个问题，把他的创造称为置信区间，回避使用概率这个词。但是鲍利及其他同行一眼就看穿了这个手法。

费歇尔也在批判者之中，不过他没有抓住这个要点。他所讨论的内容空洞又含混，而且根本不是奈曼论文里的内容。因为费歇尔根本没有完全弄清楚区间估计值的计算过程。在他

的评论里，他所指的是“信念概率”（fiducial probability），而奈曼的论文里并没有这个词汇。长久以来，费歇尔一直试图解决这个问题——怎样确定与一个参数的区间估计相关联的不确定度？费歇尔从一个很复杂的角度来解决这个问题，有点像他的似然函数。不过他很快就证明，用这种方式研究这个公式并不符合概率分布的要求。费歇尔称这个函数为“信念分布”（fiducial distribution），但他后来又违反了他自己的思路，使用了其他人在处理适当概率分布时可能会用到的相同数学方法。费歇尔所希望的结果，是从观测数据中得到参数的一组合理的值。

这也正是奈曼所得的结果，而且如果该参数为正态分布的平均数时，两个方法会得到相同的答案。据此费歇尔认为奈曼窃取了他的偏偏分布的思想，只是换了个名字而已。费歇尔对他的信念分布的研究从来没有取得进一步的发展，因为他的方法在遇到更复杂的参数（比如标准差）时就不管用了。奈曼的方法对处理任何类型的参数都是有效的。费歇尔似乎从未理解这两种方法之间的差异，直到死前他还坚持认为，奈曼的置信区间最多只是他的信念区间（fiducial intervals）概念的推广。他坚信，在碰到足够复杂的问题时，奈曼的显然是推广的方法也不会奏效——就像他自己的信念区间方法一样。

概率与置信水平

不管碰到的问题有多复杂，奈曼的方法没有失败，这也是该方法在统计分析中得到广泛应用的原因之一。奈曼置信区间中的真正问题，倒不是费歇尔所提出的那个，而是鲍利在一开始讨论时就点出来的问题，即这个方法中的概率到底指的是什么？奈曼的回答又回到了现实生活中概率的频数定义上。正如他在这篇论文里所说的（他在稍后的另一篇探讨置信区间的论文里，对这一点做了更清楚的解释），不应该从每一个结论的角度看待置信区间，而应该其视为一个过程。从长期来看，对于一直计算 95%的置信区间的统计学家来说，他们将发现，在总次数中，参数的真值将有 95%的机会落在所计算的区间内。请注意，对奈曼来说，与置信区间相联系的概率并不是我们“答对”的概率，而是统计学家使用某种方法从长期来看做出正确陈述的频率。这个数字与当前的估计值有多“准确”根本没有任何关系。

尽管奈曼定义这个概念时非常仔细，尽管许多像鲍利这样的统计学家也都非常小心，力图保持对概率概念的清晰理解并使其不被误用，但在科学领域中对置信区间的普遍应用却导致了许多草率的思维。举例来说，有人使用 95%的置信区间来表示他有“95%的把握”保证参数的真值会落在这个区间里，这是很普遍的。我们在 13 章会碰到：L·J·萨维奇和布鲁诺·德费奈蒂（Bruno de Finetti），并介绍他们对个人概率的研究，他们的研究结果证明了使用上述陈述的合理性。但是，计算某人对某一件事的把握程度，与计算一个置信区间完全是两回事。统计文献里有很多文章都谈到，根据一组相同的数据，以萨维奇和德费奈蒂的方法所推导出的参数范围，和以奈曼的方法为基础推导出的置信界限，两者之间是截然不同的。

尽管在奈曼的方法中人们对概率的涵义仍存有疑问，但是奈曼的置信界限已经成为计算区间估计值的标准方法。许多文学家计算 90%或 95%的置信界限，而且看上去好像他们有把握认为，该区间包含了参数的真值。

时至今日，已无人再谈论或在写作中涉及费歇尔的“信念分布”的话题了。该思想已随费歇尔的去世而消失。费歇尔竭力让他的思想能发挥作用，他做了大量的相当聪明而且非常重要的研究工作，其中有些研究成果已成为当今的主流，而其它部分则仍停留在费歇尔搁笔时的不成熟状态。

在费歇尔的研究过程中，他曾有好几次差点儿就建立一门统计学业的分支学科，也就是他所称的“逆概率”（inverse probability），但每次他都半途而废。逆概率的思想起源于

18 世纪的一位业余数学家雷韦朗·托马斯·贝叶斯（Reverend Thomas Bayes），贝叶斯与很多同时代的顶尖科学家都有密切的书信往来，并经常提出一些很复杂的数学问题给他们。有一天，他随意玩弄一些概率的标准数学公式，用简单的代数把其中两个式子结合在一起，竟发现一些令他很惊讶的结果。

下一章，我们来谈谈贝叶斯异论（Bayesian heresy），并且看看为什么费歇尔拒绝使用这种逆概率。

第 13 章 贝叶斯异论

从 8 世纪的早期，威尼斯共和国是地中海一带的一个主要的强权国家。在其政权鼎盛时期，威尼斯控制了大部分的亚得里亚海岸，以及克里特岛和赛浦路斯岛，同时还垄断了东方通往欧洲的商业贸易路线。威尼斯共和国由一群贵族家族所统治，这些家族之间保持着某种民主的程序。整个国家名义上的领袖是总督，从公元 697 年该共和国成立起，到 1797 年被奥地利吞并，总共有 150 余任总督，有的任期很短，只有 1 年或不到 1 年，也有的任期长达 34 年。在在的总督去世之后，该共和国会遵守一项很复杂的选举程序，他们先从贵族家族的长者当中，以抽签的方式选出一小群元老，这些被选出的元老还会再挑选一些人加入到他们之中，之后再从这一扩大的元老群中以抽签方式选出一小群人。这样的程序进行几次之后，会选出一群最后的总督候选人，总督就在这群人当中产生。

在威尼斯共和国历史的早期，每阶段的抽签都要准备一批大小相同的蜡球，有的蜡球里什么都没有，有的蜡球里面却有一张小纸条，上面写着“元老”二字。到了 17 世纪，最后几个阶段用的道具是大小完全相同的金球与银球。公元 1268 年，当多杰·拉伊涅里·泽诺（Doge Rainieri Zeno）总督去世时，在第二阶段有 30 位元老，于是准备了 30 个蜡球，其中 9 个蜡球内藏有“元老”纸条。一个小孩被带过来，他从装有蜡球的篮子中取出一个蜡球，交给第一位元老候选人，这位元老候选人就打开蜡球，看看自己是否能够成为下一阶段的元老候选人。接着，小孩从篮子中取出第二个蜡球，交给第二位元老候选人，第二位再打开蜡球，以此类推。

在小孩选出第一个蜡球前，候选人群中的每个成员被选为下个阶段元老的概率是 9/30。如果第一个蜡球是空的，剩下的候选人中每个人有 9/29 的概率成为下坠估摸元老。但如果第一个蜡球里有纸条，则其余人被选中的机会就剩下 8/29。一旦第二个蜡球被选定且被打开，则下一个人被选中成为元老的概率同样会减少或增加，是减少还是增加取决于前次的抽球结果。这样继续抽下去，直到所有的 9 个纸条都被抽出为止。而在这时，剩下的候选人下一阶段成为元老的概率就降为零。

这是条件概率的一个例子。某一特定候选人被选为下一阶段元老的概率，取决于在他的选择之前被选出的蜡球。J·M·凯恩斯曾指出，所有的概率都是条件概率。用凯恩斯所举的一个例子：从他的图书室的书架上随机地选择一本书，而选中的书是精装本的概率，也是一种条件概率，其条件取决于他的图书室里究竟有多少书，以及他怎样“随机”地选取。一个病人患小细胞肺癌的概率，是以该病人的吸烟史为条件的。对一个控制实验，检验没有处理效果这一零假设所计算出来的 P 值，是以该实验的设计为条件的。条件概率的重要方面是，某些已知事件（例如在彩票发行过程中，某一组特定数字能赢）的概率，会随前提条件的不同而不同。

在 18 世纪，为处理条件概率而导出的公式都是根据以下的思想做出的，即条件事件要发生在所研究的事件之前。但是到了 18 世纪后期，R·T·贝叶斯在摆弄条件概率的公式时，忽然有个惊人的发现，这些公式都是内部对称的！

假设有两个事件在一段时期内发生，就像先洗牌，再发出 5 张扑克牌。我们称这两个事件分别为“前事件”（the events before）和“后事件”（the events after）。以“前事件”为条件讨论“后事件”的概率是有意义的。如果牌没有洗好，当然会影响玩家得到一对 A 的概率。贝叶斯发现，我们也可以“后事件”为条件计算“前事件”发生的概率。这是没有道理的。就像玩家已经拿到一对 A 之后，再来确定整副牌里有 4 张 A 的概率。或是已知一个病人已患了肺癌，再来计算他是吸烟者的概率。或者是已经知道了有个叫 C·A·史密斯的人是唯一得到大奖的人，然后再计算州立彩票游戏公平不公平的概率。

贝叶斯把这些计算结果搁置起来，没有发表。在他死后，这些论文才被发现，而后才

被发表出来。从那里起，贝叶斯定理[16]就困扰着许多统计分析数学家。绝对不是毫无道理，贝叶斯将条件概率倒转过来反倒很有意义。当流行病学家试图想找出某种罕见医学病状的可能原因时，例如雷氏症候群（Reye’s syndrome），他们通常是利用病例控制研究方法（case-control study），在这种研究中，他们首先搜集一组患有该病症的病人，然后拿去与控制组的病人做比较，控制组的病人没有患这种疾病，但在其他方面与患有这种疾病的病人类似。于是，流行病学家在已知控制组病人已患有该疾病的条件下，计算某些先前治疗或先前条件导致该病的概率。吸烟对心脏病和肺癌都有影响，就是这样首次被发现的。镇静剂对新生儿畸形的影响，也是从这种病例控制研究中发现的。

直接应用贝叶斯定理，可以把条件概率反转过来，比这更为重要的，是使用贝叶斯定理估计分布的参数。有一种建议，可以把一项分布的参数本身看作是随机的，然后计算与这些参数相关的概率。例如，我们可能想要比较两种癌症治疗方法，并希望得到结论说“我们有 95%的把握认为使用治疗方法 A 会比使用治疗方法 B 的 5 年期存活率高”。我们只要应用贝叶斯定理一两次就可以解决这个问题。

关于“逆概率”的问题

有很多年，以这种方式使用贝叶斯定理被认为是一种不适当的作法。当用于参数时，关于概率代表什么涵义有很多质疑。毕竟皮尔逊革命（Pearsonian revolution）的整个基础在于，科学的测量结果本身不再是我们所感兴趣的问题，相反，正如 K·皮尔逊所指出的那样，我们所感兴趣的是这些测量结果的概率分布，而科学的调查研究的目的就是要估计出控制这些分布的那些参数值（固定的但却是未知的）。所以，如果这些参数被视为是随机的（而且以观测的测量结果为条件），那么这种方法就不再有这样清楚的意义了。

在 20 世纪的早些年，统计学家非常谨慎，避免使用人们所说的“逆概率”。有一次在皇家统计学会上，对费歇尔的一篇早期论文进行讨论时，就有人质疑他使用了逆概率，他坚定地为自己辩护，否认这项可怕的指控。在第一篇关于置信区间的论文里，奈曼似乎使用了逆概率的概念，但只是作为一个数学方法，用来得到一个计算结果，而在他的第二篇论文里，他证明不了不用贝叶斯定理也能得到相同的结果。到了 20 世纪 60 年代，为种方法的潜在力量与用途已开始吸引越来越多的研究者跟踪研究，这个贝叶斯异论变得越来越受尊重了。到了 20 世纪末，它已经达到了如此高的接受水平，如今在一些期刊像《统计年报》（Annals of Statistics）和《生物统计》上，几乎半数以上的文章现在都使用贝叶斯方法。不过，贝叶斯方法的应用仍然会经常遭到质疑，尤其是在医学领域。

在解释贝叶斯异论时碰到的一个困难是，目前有好几种不同的分析方法，而这些方法的应用又至少有两种完全不同的哲学基础。长期以来，看上去好像完全不同的思想却经常贴着相同的标签——贝叶斯。后面我将说明贝叶斯异论的两个种理论：贝叶斯层次模型（Bayesian hierarchal model）和个人概率（personal probability）。

贝叶斯层次模型

20 世纪 70 年代早期，由于弗雷德里克·莫斯特勒（Frederidck Mosteller）和大卫·华莱士（David Wallace）早期的工作和贡献，原文分析的统计方法有了很大的进展，他们俩人曾运用统计方法来判定《联邦主义论文集》（Federalist）中一些匿名文章的作者。自 1787

[16] 施蒂格勒的误称定律在贝叶斯定理（Bayes’s theorem）这个名字上得到完全显现。贝叶斯绝对不是首先发现这种条件概率对称性的人。贝努里似乎已经注意到它，棣莫弗也曾提到它。但却由贝叶斯独享盛名（若以贝叶斯不愿意发表其成果的心态来看，我们或许可以说贝叶斯已经受到了谴责）。

年，在纽约州带头鼓动通过新的美国宪法期间，詹姆士·麦迪逊（James Madison）、亚力山大·汉密尔顿（Alexander Hamilton）和约翰·杰伊（John Jay）写了大约 70 篇文章，支持通过宪法。但这些文章都是匿名发表的。19 世纪初，汉密尔顿与麦迪逊两人开始确认这些两个人都声称有著作权的论文，其中有 12 篇文章他们都认为是自己写的[17]。

在用统计方法对这些署名有争议性的文章进行分析时，莫斯特勒与华莱士找出了几百个无"特定内容"的英文词汇，如"if"、"when"、"because"、"over"、"whilst"、"as"、"and"等。这些字在句子里只有语法上的意义，本身并没有什么特定的含义，这些字的使用主要取决于作者的语言使用习惯。在这上百个没什么特定含义的字里，他们发现，大约有 30 个字在这两位作者的其他著作中使用频率不同。

例如，麦迪逊使用"upon"这个字的频率，是每千字平均 0.23 次，但汉密尔顿对这个字的使用频率很高，平均每千字高达 3.24 次（在 12 篇署名有争议的文章里，有 11 篇根本没有用"upon"这个字，而在剩下的那一篇文章中，平均每千字就出现 1.1 次）。这些平均的频率并不是描述一千字中任何特定组合。这些数值本身并不是整数，这就意味着这些频率并不是在描述任意一个观测的文字序列。这些数值其实是两位不同作者在写作时用字分布的其中一个参数的估计值。

对于某篇文章著作权的争议，所要解决的问题是：这些文章中用词的分布形态，是来自与麦迪逊相联的概率分布呢？还是来自与汉密尔顿相联的概率分布？这些分布各有各有参数，其中能够定义出各自作品的特定参数各不相同。参数值只能根据他们的论文来估计，而且这些估计可能是错的。因此，要想区分哪个分布可应用在一篇署名有争议的文章上，充满了这种不确定性。

估计这种不确定性水平的一种方法是，这两个人的分布参数的确切值，是来自于描述 18 世纪晚期所有北美洲有教养的人用英文写作时用字习惯的参数分布。例如，汉密尔顿每千字中用到"in"这个字 24 次，麦迪逊则是每千字用 23 次，而同时代的其他作家，使用"in"这个字的频率在每千字 22 至 25 次之间。

由于受到当时和当地一般用字分布形态的制约，每个人分布的参数是随机的，并且具有一个概率分布。这样一来，制约汉密尔顿和玫迪逊使用这些无特定含义的字的参数本身也有参数，我们可以称之为"超参数"（hyper-parameter）。根据当时和当地其他作者发表的文章来分析，我们就能估计出这些超参数。

英语语言总是随着时间和地域的变化而变化。例如在 20 世纪的英语文学里，使用 in 的频率通常是每千字少于 20 次，这表明从汉密尔顿和麦迪逊的时代到现在的 200 多年里，英语的用字型态已经稍微有所转变。我们可以把这些定义 18 世纪北美用字习惯参数分布的越参数，看作是它们本身也有一个相对于所有时间与空间的概率分布。因此，除了用 18 世纪的北美作品，我们还可以搜集其它地区和其它时期的英语文献，来估计这些超参数的参数，我们可以称这些参数为"超－超参数"（hyper-hyperparameter）。

通过重复使用贝叶斯定理，我们就能决定这些参数的分布，然后再决定这些超参数的分布。从原则上来说，我们可以用超－超－超参数求出超－超参数的分布，进而把这种层次分析引向深入，依次类推。但在我们的例子里，显然没有必要进一步分析，以免增添更多的不确定性。利用超参数与超－超参数的估计值，莫斯特勒与华莱士就能算出与下面这个陈述有关的概率：是麦迪逊还是汉密尔顿写了这篇文章。

自 20 世纪 80 年代早期以来，贝叶斯层次模型已经成功地解决了许多工程上和生物学上的难题。比如，一些数据看上去似乎是来自于两个或两个以上不同的分布，这个问题就属于这类难题。分析家可以建议，有一个未观测到的变量存在，而这个变量可以定义已知的一

[17] 实际上，做此声称的只有麦迪逊。在汉密尔顿去世 3 年后，他的朋友把据称是他写的论文整理发表后，麦迪逊就提出，其中 12 篇是自己写的。

个观测结果究竟来自于哪个分布。这个差别标识本身是个参数。但它还有一个概率分布（含有超参数），这个概率分布可以纳入到似然函数当中来进行分析。莱尔德和韦尔的 EM 演算法特别适合于解决这类问题。

统计文献中对贝叶斯方法的广泛使用充满了混淆与争议。大家可以提出得出不同结果的不同方法，但却没有明确的标准来决定哪个是对的。通常，保守肖像统计学家反对使用贝叶斯定理，而贝叶斯学派的人彼此对他们模型的细节看法也不一致。这种混乱的状况亟需另一个像费歇尔这样的天才出现，找出一个统一的原则来解决这些争议。当我们进入 21 世纪的时候，还没有这样的天才出现。因此，相关的问题还是像在 200 多年前的贝叶斯时代一样，令人困惑。

个人概率

另外一种贝叶斯方法其基础看上去要坚实得多。这就是个人概率（personal probability）的概念。个人概率的意思自从 17 世纪贝努里一开始研究概率时就已经产生了。实际上，概率（probability）这个英文字创造的初衷，就是用来处理主观不确定性的。

L・J・萨维奇和布鲁诺・德费奈蒂在 20 世纪 60 年代和 70 年代，推导出了个人概率背后的许多数学模式。我在 20 世纪 60 年代末期曾参加一场在北卡罗来纳大学举办的统计学会议，会上萨维奇在演讲中曾阐述他的一部分想法。萨维奇认为，世界上并没有“已被证明的科学事实”这样的事情。有的只是一些陈述，而那些自认为是科学家的人对这些陈述持有很高的赞成概率。他举例说，在场听他演讲的人对“地球是圆的”这项陈述一定持有很高的认同概率，但若我们有机会对全世界的人做一次普查，则我们很可能发现在中国中部的许多农民对上述陈述持有很低的概率。讲到这里的时候，萨维奇不得不被迫停下来，因为校园晨一群学生正在会堂外游行通过。他们还高喊着口号“停止上课！罢课！罢课！停止上课！”这些学生在要求全校的学生罢课，以抗议越南战争。等到他们走远，四周又恢复平静，萨维奇才看看窗外，然后说：“看来，我们可能是认为地球是圆的人中的最后一代。

个人概率有许多不同的版本。其中一个极端是萨维奇－德费奈蒂的方法，该方法认为每个人都有其自己独特的一套概率。而另一个极端则是凯恩斯的观点，他认为概率是一种信仰程度（the degree of belief），这种信仰是一个在特定的文化环境中一个有教养的人可能期望持有的信念。按照凯恩斯的观点，一个特定文化环境中的所有人（萨维奇所说的科学家或中国中部的农民）对某一特定的陈述，会持有一个一般的概率水平。由于这个概率水平取决于文化和时间，因此从某种绝对的意义上为说，很有可能这个适当的概率水平是错的。

萨维奇和德费奈蒂则主张每个人都有自己特定的一套个人概率，他们还描述怎样运用一种叫做“标准赌博”（standard gamble）的技巧把这种人人概率求出来。为了让整个文化中的人能共享既定的一套概率，凯恩斯不得不弱化相关的数学定义，概率不再是一个精确的数字（例如 67%），而是一种将想法排序的方法（例如，明天可能下雨的概率大于可能下雪的概率）。

不管个人概率的概念是如何被准确定义的，贝叶斯定理在个人概率中的应用方式，看上去与大多数的想法相吻合。贝叶斯方法一开始是假设在一个人的头脑中有一组先验概率（a prior set of probabilities），接下来这个人经过观测或实验产生了数据，然后再拿这组数据来修正先验概率（prior probability），生成一组后验概率（a posterior set of probabilities）：

先验概率　→　数据　→　后验概率

假设这个人想确定是否所有的大乌鸦都是黑的。她首先存有一些关于“这个陈述是真的”概率的先验知识。例如，起初她可能对大乌鸦一无所知，对“所有大乌鸦都是黑的”这句话半信半疑，相信比例是 50：50。数据则包括她对大乌鸦的观测。假如她看到了一只大

乌鸦，而且这只大乌鸦是黑色的，她的后验概率就会增加。因此下一次她再观测大乌鸦时，她的新的先验概率（也就是上一次的后验概率）就会大于 50%，如果她继续观测大乌鸦而且都是黑的，这个概率还会继续上升。

另一方面，一个人也有可能在进行观测之前就已经带着非常强的事前主见，其程度非常强，需要有很大量的数据才能改变这个事前主见。在 20 世纪 80 年代，美国宾夕法尼亚州的三里岛核电厂发生了近乎是灾难性的事故。反应炉的操作员面对一个很大的操作盘，通过上面的各种仪表和指示灯来了解反应炉的运转情况。这些指示灯当中有一些是警告灯，其中有的出过问题，以前曾经发出过假的警告。当时操作员有个事先的成见，当他们看见任何一个新的警告灯亮时，总是认为它是假的信号。结果，即使当警告灯的型态及相关的指示器都一致显示反应炉的水位过低时，他们仍然置之不理。他们的先验概率太强了，以至于新的数据也无法使后验概率产生多大的改变。

假定只有两种可能性，就像前面署名有争议的联邦主义论文的例子：它不是麦迪逊写的就是汉密尔顿写的。于是，在应用了贝叶斯定理之后，就会得到了一个先验胜率（prior odds）与后验胜率（posterior odds）之间的简单关系，这里的数据可以归纳成一种称为"贝叶斯因子"（Bayes factor）的东西。这是一种根本不用参考先验胜率来刻画数据的一种数学计算。有了这个计算工具，分析家就可以告诉读者，插入任何他想要的先验胜率，乘以计算出来的贝叶斯因子，再计算后验胜率。莫斯特勒与华莱士对 12 篇署名有争议的文章，每篇都是这样处理的。

此外，他们对文章里的那些无特定含义的字出现的频率，还进行了两种非贝叶斯分析。这样他们有了四种方法来判断有争议文章的作者：层次贝叶斯模型，计算的贝叶斯因子，以及两个非贝叶斯分析方法。结果如何呢？所有 12 篇文章都压倒性地指向麦迪逊。实际上，如果使用计算的贝叶斯因子，那么对某几篇文章来说，读者认为是汉密尔顿写的先验胜率可能要大于 100000：1 才有办法让后验胜率为 50：50。

第 14 章 数学界的莫扎特

在 20 世纪统计学方法的发展历程中，费歇尔并不是唯一的天才。俄国数学家安德烈 • N • 柯尔莫哥洛夫（Andrei N. Kolmogorov）（比费歇尔年轻 13 岁，1987 年以 85 岁高龄过世），在数理统计与概率理论方面留下了很多不朽的成就。他的成就虽然是以费歇尔的一些研究成果为基础的，但柯尔莫哥洛夫的成就在数学深度与细节上都超越了费歇尔。

不过，就像他的成就对科学的贡献非常重要一样，柯尔莫哥洛夫对所有认识他的人也颇具影响力。他的学生艾伯特 • N • 谢耶夫（Albert N. Shiryaev）在 1991 年写道：

> A · N · 柯尔莫哥洛夫属于那种很少数、你一接触就知道他与众不同的人，他很伟大、很杰出，感觉像个奇才。他的一切都和别人不一样：他的一生，他的中学和大学生活，他在数学……气象学、流体力学、历史、语言学、教育学等领域的开创性发现。他的兴趣异常广泛，包括音乐、建筑、诗歌及旅行。他的博学多闻也是罕见的。看上去好像他对任何事都有很高深的见解……。任何人只要和他见过面，只要与他简单交流，便会感觉他是那样的非常寻常。人们感觉到，他是那种具有连续深度心智活动的人。

柯尔莫哥洛夫生于 1903 年，那年他的母亲正从克里米亚（Crimea）返回家乡，她的家乡在俄国南部托诺西纳（Tunoshna）的乡村，在旅行途中生下了柯尔莫哥洛夫。有一位传记作家很精确地写到："柯尔莫哥洛夫是个非婚生的儿子。"他的母亲玛丽亚 •雅科夫列夫娜 •柯尔莫哥洛夫（Mariya Yakovlevna Kolmogorov）在怀孕的后期被其男朋友抛弃，只得回家待产，不料阵痛提早发作，她只好在中途的坦波夫（Tambov）镇下了火车，在那儿生下了小孩。不幸的是，她自己却因难产死于这个陌生的小镇，只有她的初生婴儿回到了故乡托诺西纳。后来是他妈妈的几个未婚姊妹抚养了他，其中的薇拉 •雅科夫列夫娜（Vera Yakovlevna），后来变成了他的养母。阿姨们为年轻的安德烈和他同龄的孩子在村子里办了一个小学校。她们甚至在家里印刷了一份小刊物，叫做《春燕》（Spring Swallows），他的第一篇作文就发表在上面。在他 5 岁的时候，他提出了他的第一个数学发现（也发表在《春燕》上）。他发现最小的 k 个奇数和和正好等于 k 的平方。随着他慢慢长大，他常拿一些问题问同学，这些问题与它们的答案也发表在《春燕》上。其中一个问题是这样的：缝一个四孔的钮扣，有多少种缝法？

到了 14 岁，柯尔莫哥洛夫从百科全书上学到一些高等数学，并且补充了其中没有证明的部分。在念高中的时候，他的一系列永动机的制造计划，考倒了年轻的物理老师。因为计划制定得太精巧了，连老师都不能发现其中的错误（柯尔莫哥洛夫把这些错误很小心地隐藏起来）。后来，他决定提早一年参加毕业考试。于是就正式向老师提出请求，老师要他午饭后回来听消息，然后他就出去散步了。等他回来的时候，学校考试委员会决定不必经过考试就发了证书给他。他后来对谢耶夫表示，这件事是他一生中最令人失望的事情之一，本来他希望迎接智力的挑战。

1920 年，年仅 17 岁的柯尔莫哥洛夫来到莫斯科念大学。他注册读数学第，但到很多别的科系去听课，如冶金学，另外他还参加一个研究俄国历史的专题研讨会。作为研讨会的一部分内容，他报告了他的第一篇等待发表的研究论文，内容是分析 15 到 16 世纪时诺夫哥罗德（Novgorod）地区土地占有情况。他的教授批评这篇论文，认为柯尔莫哥洛夫没有提供足够的证据。几年后，有个考古队在该地区探险，证实了柯尔莫哥洛夫的猜测。

作为莫斯科国立大学的学生，他到中学兼职做教员，还参加了许多课外活动。后来他继续在莫斯科大学读数学专业的研究生。数学系要求学生修 14 门基础课程，而对于每门课程，学生可以选择或是参加期末考试，或是提交一篇具独创性的论文。很少有学生尝试写出一篇以上的论文柯尔莫哥洛夫从没参加过考试，而是写了 14 篇具独创性的精彩论文。他后来回忆说，"其中一篇的结果其实是错的，但我只是在后来才意识到。"

柯尔莫哥洛夫这位才华横溢的数学家得到西方科学家的赏识，是通过他在德国出版的一系列精彩的文章及一些德文书籍实现的。在 20 世纪 30 年代，俄国当局甚至还允许他去参加一些在德国和斯堪的那维亚举行的数学研讨会。不过在第二次世界大战期间以及战后，柯尔莫哥洛夫这个伟大的人物却消失在斯大林的铁幕后面。1938 年，他发表了一篇论文，这篇论文建立了平滑和预测平稳随机过程的基本定理（这项研究在本章后部分还将做介绍）。诺伯特・维纳（Norbert Wiener）对于战争的状态给出了一个有趣的评论，维纳当时正在麻省理工学院（Massachusetts Institute of Technology），在战争期间和战后，他致力于将这些方法应用于军事问题。维纳的研究结果被认为对美国的冷战非常重要，以至于被宣布为最高级的机密。但是维纳坚持认为，他的所有研究结果都可以从柯尔莫哥洛夫早期的那篇论文中推导出来。在二次大战期间，柯尔莫哥洛夫忙于研究如何将该理论应用于苏联的战争中。柯尔莫哥洛夫一直谦逊地评价自己的学术成就，他认为这些基本思想应该归功于费歇尔，因为费歇尔在他的遗传学的研究中使用了类似的方法。

柯尔莫哥洛夫其人其事

1953 年斯大林去世后，政治上处处怀疑的铁环开始松动。于是柯尔莫哥洛夫这个人又开始露面，参加一些国际学术会议，同时在俄国也组织一些学术会议。国际上的数学界开始认识他。他是一个热心、友善、开明、幽默的人，同时知识渊博，喜爱教学。他那敏锐的大脑对他的所见所疗总是不停地在思考。我手头有一张 1963 年柯尔莫哥洛夫在第比利斯（Tbilisi）听英国统计学家大卫・肯德尔（David Kendall）讲座时的照片，柯尔莫哥洛夫的眼镜搭在他的鼻尖上，他身体前倾，热切地跟踪讨论。你可以感觉到一种鲜明的个性，感染着坐在他周围的人。

柯尔莫哥洛夫最喜爱的一些活动是给莫斯科的一些有天赋的孩子讲课并组织课堂活动，他非常乐于将孩子们引入到文学和音乐的知识领域。他带孩子们远足和探险，他认为每个孩子都应该有一个“完整个性的宽广而自然的发展空间”。大卫・肯德尔曾写道：“这些孩子将来是不是都成为数学家，这并不是他所关心的。不管孩子们最终从事什么职业，只要他们的远见仍然宽阔，只要他们的好奇心并没有被遏制，他就会感到满意。”

柯尔莫哥洛夫在 1942 年与安娜・德米特里耶夫那・叶戈罗娃（Anna Dmitrievna Egorova）结婚。他们恩爱美满的婚姻一直延续到他们 80 多岁。他是一位狂热的徒步旅行和滑雪爱好者，在他 70 多岁的时候，还带领年轻人远足攀登他所喜欢的山脉，讨论数学、文学、音乐和普通的生活问题。1971 年，他加入了一个科学探险队，在德米特里・门捷列夫（Dmitri Mendeleev）科学考察探险船上探索海洋的奥秘。他的同辈不断地对他所感兴趣的事物和他所拥有的知识感到惊奇。在他会见约翰・保罗教皇二世（Pope John Paul Ⅱ）时，他与这个爱好运动的教皇讨论滑雪，并指出，在 19 世纪，胖的教皇与瘦的教皇交替出现，并且还指出约翰・保罗教皇二世是第 264 任教皇。看上去他的研究兴趣之一是罗马天主教的历史。他曾经做关于俄国诗歌的统计分析方面的讲座，他还能记住并大段大段地背诵普希金（Pushkin）的诗歌。

1953 年，莫斯科国立大学组织了一次大型活动，庆祝柯尔莫哥洛夫的 50 生日。作为该活动的一位演讲人，该校退休的名誉教授帕维尔・亚历山大德罗夫（Pavel Aleksandrov）曾讲到：

> 柯尔莫哥洛夫属于这样一类数学家，他们在任何一个领域中的每一项研究都会引领出一种全新的评价。在这些年，我们很难找到一个像他这样的数学家，不但兴趣广泛，而且对数学界深具影响力，……哈代（Hardy，一位著名的英国数学家）认为他是三解级数的专家，而冯・卡曼（von Karman，一位二次大战后的德国物理学家）则认为它是

机械学专家。格德尔（Gödel，一位数学哲学理论学家）曾说，天才的特质是永远保持着童心。所谓的童心有许多特质，感到兴奋是其中之一。对数学感到兴奋是柯尔莫哥洛夫作为天才的一个印证。除此之外，柯尔莫哥洛夫对事物的兴奋，还展现在他具创造性的研究成果中，在他为《俄国百科大辞典》（Large Soviet Encyclopedia）写的许多文章里，在他所开发的博士项目中。这些都只是他的一个方面，而他的另外一面，则是他专心致志的做事态度。

他这种专心致志的做事态度其结果是什么呢？要列出柯尔莫哥洛夫在数学、物理、生物与哲学领域中有哪些重要贡献，倒不如列出他在这些领域里的哪一方面没有多大贡献，后者比前者容易得多。1941 年，他建立了研究湍流的现代数学理论方法。1954 年，他在检验行星间的重力交互作用时，发现了一种模拟方法，可用来描述其中的“不可积分”性，这正是百年来数学分析所面临的一个挑战。

柯尔莫哥洛夫在数理统计方面所做的工作

对于统计学的革命，柯尔莫哥洛夫解决了两项最迫切的理论问题。在他去世之前同，他几乎解出了困扰统计方法核心的一个很深奥的数学哲学问题。这两个迫切的问题是：

1. 概率的真正数学基础是什么？
2. 面对像地震过后的余震（或地下核弹试爆）这类长时间搜集上来的数据时，我们可以做些什么？

当柯尔莫哥洛夫开始研究第一个问题时，概率在理论数学家的眼里名声并不太好。这是因为，很多人认为创建于 18 世纪的计算概率的数学技巧，只不过是比较聪明的计数法而已。（例如，从一副标准的扑克牌中，抽取 3 组牌，每组 5 张，可以有多少种发牌法只会让其中一位参与者成为赢家？）这些聪明的计数方法看上去似乎没有一个单一的基础理论结构，好像都是为了满足某项特殊需求而创造出来的特定做法。

对大部分的人来说，有个能解决问题的方法就够了，但对 19 世纪末、20 世纪初的数学家来说这是不够的，他们需要一个坚实而严密的基础理论，以确保得到的这些解中不会有错误。18 世纪数学家们所使用的这些特定方法虽然有用，但如果应用错了也会产生很难应付的悖论。因此，20 世纪初期数学的主要工作就是把这些特定方法放在一个坚实而严密的数学基础上。亨利•勒贝格（就是让奈曼印象非常深刻的那位很有数学见地的勒贝格，但后来奈曼真的与他见面时，却觉得他粗鲁而没礼貌）的研究工作之所以这么重要，就是因为他把微积分的特殊方法建立在一个坚实的基础上。只要概率理论还停留在 17 和 18 世纪那种不完整的阶段，20 世纪的数学迷朦就会认为概率理论是一种没多大价值的东西（许多统计方法也会遭此轻视）。

柯尔莫哥洛夫思考了概率计算的本质之后，最后终于发现，求一个事件的概率完全就像求一个不规则形状的面积。他把新产生的数学测试理论应用到概率的计算上。有了这些工具，他就能定出一套公理，再用这些公理建构出整个概率理论。这就是柯尔莫哥洛夫的“概率论的公理化”（axiomization of probability theory），至今仍是学校中讲授概率论时采用的唯一方法。这种方法永久性地解决了有关概率计算有效性的所有问题。

解决了概率理论的问题之后，柯尔莫哥洛夫开始攻关另一个有关统计方法的主要问题（与此同时，他还要教那些天才的儿童，组织研讨会，管理数学系，解决有关机械学与天文学的问题，以及如何让生活过得既充实又精彩）。为了使统计计算变得可行，费歇尔以及其他的统计学们家都假设所有的数据都是独立的。他们把一系列的测量结果看成像是掷骰子得来的。因为骰子没有记忆，不会记得它们上次出现的点数，所以每次新出现的点数都与先前出现的点数完全独立。

大部分数据并不是彼此独立的。费歇尔在《研究工作者的统计方法》一书中所举的第一个例子，是他的新生儿子每周的体重。显然，若小孩在一星期内增加很多体重，下一周的数据当然会反映这种结果；如果小孩此周生了病，体重没有增加，下周的体重数据也会把这个结果反映出来。在现实生活中，一个长时间搜集上来的数据序列很难被认为是真正独立的。

费歇尔在他的《作物收成变动研究》这一著作的第三篇中（也就是 H·费尔菲尔德·史密斯教授介绍给我的那篇重量级论文），记录了连续几年的小麦收成量和那几年每日的降雨量。随时间所搜集得来的数据并不是独立的，他通过创建一组很复杂的参数来应对这一难题。他找到了一些有限的解，但这些解所根据的简化假设可能并不成立。费歇尔无法再进一步解决这个问题，也没有人继续从事他这项未完成的研究。

当然，我们说的没有人，是指在柯尔莫哥洛夫出现之前。柯尔莫哥洛夫把随时间搜集得来的前后相联的这一数值序列，称作“随机过程”（stochastic process）。他的许多篇先驱性论文（正好在二次世界大战爆发前发表）为美国的 N·维纳、英国的乔治·博克斯（George Box）以及他自己在俄国的学生进行更深入的研究奠定了基础。由于有了柯尔莫哥洛夫的思想，现在我们已经能够对那些随时间搜集上来的纪录时行检查分析，而且可以得出很专门的结论。我们可以利用加州海岸的海浪数据来定位印度洋上的风暴；无线电波望远镜能区分不同来源的无线电波（或许有一天甚至还能接收到其它星球上高等生物发出的信息）；我们有可能分辨一组震波纪录究竟是地下核弹试爆引起的，还是天然的地震引起的。在工程学的期刊上，许多文章所采用的方法都是根据柯尔莫哥洛夫对随机过程的研究成果而发展出来的。

现实生活中概率的意义是什么？

在生前的最后几年，柯尔莫哥洛夫攻关一个更困难的问题，这个问题不公是个数学问题，而且还是个哲学问题。到他去世的时候，这个问题还没有完全获得解决。不过，一代数学家已经在认真思考如何接续他的思路进行研究。在我写这本书的时候，这个问题还没有解决。不过，正如我在最后一章将要指出的，如果这个问题一直无法解决，那么对科学来说，统计方法的整个体系就会被它自己的前后不一致所搞垮。

柯尔莫哥洛夫研究的最后一个问题是：在现实生活中，概率的意义是什么？他已经为概率提出了一个令人满意的数学理论。这意味着，概率的所有定理和方法都是内部自身前后一致的。科学的统计模型则跳出了纯数学领域，把这些定理应用在现实问题上。为了做到这一点，柯尔莫哥洛夫为概率理论所提出的抽象数学模型，必须找到与现实生活某些方面的对应关系。实际上已有上百种方法想解决这个问题，每一种方法对概率在现实生活中的意义都提出了不同的解释，但每种方法都受到了批判。这个问题非常重要，因为如何解释统计分析的数学结论的涵义，取决于你如何在这些公理与现实生活中的情况之间找到对应的关系。

在柯尔莫哥洛夫的概率理论的公理化过程中，我们假设存在一个抽象空间，空间里的元素称为“事件”（event）。该空间中事件的集合，可以像我们测量门廊的地板面积或电冰箱的体积一样进行测量。如果对抽象的事件空间的测量满足某些公理，则称该空间为概率空间（probability space）。为了在现实生活中应用概率理论，我们得找到这个事件空间，而且要非常明确具体，这样我们才能实际计算出该空间概率的测试值。当一个实验科学家使用统计模型来分析实验的结果时，这个空间是什么？威廉·西利·戈塞特认为这个空间是实验的所有可能结果的集合，但他无法证明应该怎样计算与该空间有关的概率。除非我们能够确定出柯尔莫哥洛夫的抽象空间，否则由统计分析得到的概率陈述会有很多不同的意义，有些意义还可能互相矛盾。

例如，假设我们进行一项临床实验，以检验一种艾滋病新疗法的功效。假定统计分析显示，旧的疗法和新的疗法之间的功效差异是显著的。那么这是否意味着，医学界可以确信这

一新的疗法能治愈下一个艾滋病病人呢？或者是否意味着，这个新疗法对一定百分比的艾滋病病人有效？或者仅仅是表示，只有对实验中经过高度筛选的这群艾滋病病人，新的疗法才会有效？

要找出概率的现实意义，通常可以通过柯尔莫哥洛夫的抽象概率空间给出现实的解释来实现。柯尔莫哥洛夫用的则是另外一种方式。他结合了热力学第二定律、K·皮尔逊的早期研究，以及一些美国数学家为了找出信息的数学理论所进行的研究尝试，还有保罗·利维对大数定律的研究，然后他从 1965 年开始，陆续撰写了一系列的论文，撇开了有关的公理和他自己对这一数学问题的解，而把概率视为……

1987 年 10 月 20 日，柯尔莫哥洛夫去世。而在他逝世前最后那几年，他依然活力十足，具有独创性的观念仍源源不绝地涌出——至今仍无人能拣起他留下来的线索。

苏联统计学界的失败

虽然柯尔莫哥洛夫和他的学生在概率和统计的数学理论上有重大的贡献，但苏联从这场统计革命中却获益很少。为什么会如此？这个问题本身就提供了一个案例，说明当一个政府对所有的问题都知道其“正确”答案时，会发生什么后果。

在沙皇统治时代的末期以及俄国大革命开始的这段期间，俄国的统计学界相当活跃。俄国数学家在英国和欧洲发表的论文，被国际学术界广泛知晓。俄国数学家与农业学家的论文常发表在《生物统计》期刊上。具有革命精神的俄国政府设立了一个中央统计局，并且在各个苏维埃共和国里也设置了类似功能的地方统计局。中央统计局进行了一份报导统计学术活动的期刊《统计学通报》（Vestnik statistiki － herald，1994 后改名为《统计学研究》，即 Voprosy statistiki － statistical studies——译者注），上面有很多英文与德文期刊的论文摘要。在 1924 年年末，《统计学通报》上发表了一篇论述统计设计如何应用在农业研究上的文章。

随着 20 世纪 30 年代斯大林肃反运动的到来，所谓正宗的共产主义理论也渗透到学术界各个领域。在一些所谓的共产主义理论家看来，统计学是社会科学的一个分支。所有的社会科学都应服从于中央计划。随机变量的数学概念是统计方法的核心，但由于随机变量（random variable）译成俄文时，译成了“偶发数量”（accidental magnitude），所以对中央计划者和理论家来说，这种概念显然是一种冒犯。在前苏联，所有的工业与社会活动，都是计划出来的，没有什么事是偶然发生的。偶发数量可能描述资本主义经济中所观察到的事情，但绝不是在俄国。因此，数理统计的应用研究很快就受到压制。在 1956 年的《数理统计年报》（The Annals of Mathematical Statistics）中，S·S·扎尔科维克（S. S. Zarkovic）写了一篇回顾苏联时期统计发展史的文章，里面就很委婉地讲到：

> 随后几年，在俄国的统计学发展过程中，政治考虑成为愈来愈显要的因素，这便导致了在统计实践活动中理论应用的逐渐消失。到了 20 世纪 30 年代末期，《统计学通报》停止刊登用数学处理统计问题的论文。到了 20 世纪 30 年代结束时，这方面的论文完全销声匿迹，而且从此没再出现。这种趋势的结果是，统计学家完全放弃了应用，躲回到大学校园和其他研究机构中，以其他学科的名义从事统计研究。柯尔莫哥洛夫、N·V·斯米尔诺夫（N. V. Smirnov）、V·I·罗曼诺夫斯基（V. I. Romanovsky）以及其他很多人，都正式地离开统计学，变成数学家了。一个很有趣的例子是 E·斯卢茨基（E. Slutsky），他本来是世界知名的计量经济学大师，结果连他也放弃统计学，改行去做天文学研究……。依照官方的观点，统计学变成了为政府制定国家经济计划的工具，当然它是一种社会科学，或换句话说，是一种阶级科学。其中的大数定律、随机离差思想，以及其它任何属于统计学的数学理论，都被当成是错误通论的构成元素，而遭到清除。

不只是官方的观点制约了统计学的发展。斯大林依赖一个大言不惭的生物学业家特罗菲姆·D·李森科（Trofim D. Lysenko），他拒绝接受遗传学的基因理论，声称动植物的遗传特征可以由环境来塑，毋需藉由遗传。那些想遵行费歇尔的成果以数学方式研究遗传学的生物学家都受到排斥，有些甚至入狱。当教条的理论降临苏联统计学界时，由中央统计局和它的下属统计局报出来的数据，也越来越受质疑。在中央计划之下，乌克兰与白俄罗斯共和国的肥沃农田，都变成泥泞的荒地，一大堆粗制滥造的机械成品根本不好用，支离厂矿的消费品由工厂流出来，也根本派不上用场。苏联甚至连填饱老百姓的肚皮都存在困难。唯一有效进行的经济活动是黑市交易。然而，中央政府依然捏造出虚假、乐观的统计数字，真实的经济活动水平被许许多多的经济增长率的比率指标所掩饰了。

此时，一些美国数学家，像诺伯特·维纳，则开始利用柯尔莫哥洛夫和亚力山大·亚·赫因强（Alexander Ya Khintchine）所提出的随机过程定理，强化美国的国防事务，而美国国家标准局的沃尔特·休哈特（Walter Shewhart）与其他人，则向美国工业界展示如何运用统计方法来控制产品投师。此外，美国、欧洲及一些亚洲地区的农场，作物的产量都在飞速提高。相反，苏联的工厂仍在生产一些没有用的东西，他们的农业依然无法解决人民的温饱问题。

直到 20 世纪 50 年代，尼基塔·赫鲁晓夫（Nikita Khrushchev）开始掌权，官方理论的控制开始放松，开始尝试把统计方法应用在工业与农业上。不过，官方的“统计”仍然是充满了假的数字与精心制作的模糊内容，而尽全力试图出版的应用统计学期刊，结果也只是不定期地出了几期而已。一直到 20 世纪 90 年代末期，苏联政府与它的中央计划经济制度完全解体，俄罗斯工业界才有机会大量采用现代统计模型。

也许这件事给大家都上了宝贵的一课。

第 15 章 “小人物”之见解

弗洛伦斯·南丁格尔（florence Nightingale）是英国维多利亚时期的传奇人物。与她打交道的国会议员和军事效仿视她为一个令人头疼的人物。一般人只把她看作是护士这个行业的创始人，一个温文尔雅、具有自我牺牲精神照料病人的护士。其实，是个很有使命感的女人，同时她也是一位自修成功的统计学家。

南丁格尔的一个使命是，强迫英国军方在战地开设医院，为战场上的士兵提供护理与医疗照顾。为了支持她自己的主张，她曾埋头于研究堆积如山的军事档案。后来，她带着一系列令人瞩目的资料与图表出现在皇家委员会面前。在这些资料和图表中，她指出在克里米亚战役（Crimean War）期间，英军死亡的主要原因是在战场外染上疾病，以及战场上受伤之后没有得到及时的照料所致。为了展示她的相关数据与资料，她还发明了饼图（pid chart）。和这些愚钝而又不学无术的军事将领打交道，南丁格尔感到很疲惫，于是她就会躲到艾文顿（Ivington）小村去住上一段时间，在那里，她总是会得到她的好朋友大卫一家人的欢迎。当年轻的大卫夫妇喜获千金时，还用她的名字为女儿命名，取名叫弗洛伦斯·南丁格尔·大卫（Florence Nightingale David）。南丁格尔的充沛精力和创造精神似乎也传给了这位同名的女孩（她一生以 F·N·大卫的名字出版了 10 本书，在科学期刊上发表了一百多篇论文）。F·N·大卫在 1909 年出生，5 岁的时候，第一次世界大战的爆发中断了她受教育的正常进程。由于住在偏僻的小乡村，大卫一开始接受的教育是当地牧师办的私人学堂。这位牧师对这个小弗洛伦斯·南丁格尔·大卫的教育有一些奇特的想法。他注意到这个小女孩已经学过一些自述知识，因此就开始教她代数。他发觉她已经学过英文，因此就开始教她拉丁文和希腊文。到她 10 岁的时候，大卫才转到普通学校接受教育。

到了大卫该上学的年龄时，听到大卫想要读伦敦的大学学院（University College, London），她的母亲大吃一惊。这个大学学院是英国哲学家杰里米·边沁（Jeremy Bentham）创办的（边沁的遗体经过弄干保存，如今还穿着正式的衣服展示在学院的回廊上）。这个学校是为“野孩子、异教徒，及不愿信奉三十九条教规（即英国国教基本教义——译者注）的人”而设立的，因为在该校创办之前，进入英国所有大学的教师和学生都必须信奉英国国教。就在大卫准备进大学的时候，大学学院还是不信奉英国国教的新教徒的温床。“那时，母亲对我要到伦敦大学学院念书……总觉得不光彩、不正当，诸如此类。”因此，她最后进了伦敦的贝德福德女子学院（Bedford College）。

很久以后在一场录音谈话里，她对哈佛公共卫生学院的纳恩·莱尔德（Nan Laird）教授透露，“我非常不喜欢贝德福德学院，但我倒是很喜欢每晚到剧院看戏。如果你是学生，你就可以花 6 便士到维多利亚剧院看一场戏……我当时过得非常快乐。”她接着说，在学校里，“有 3 年时间我只学数学，其它什么都没学，但我很不喜欢这样。我甚至不太喜欢学校里的人，可能当时我很叛逆吧。不过我并不怀念那段大学岁月。”

她在学校里学了这么多的数学，毕业之后能用来干什么呢？她想当个保险精算师，但当时这个行业只招男性。有人建议她去找大学学院中一位叫 K·皮尔逊的教员，该教员研究的事情可能与精算或此类事情有关。于是她就来到大学学院，“我直接就去找了 K·皮尔逊。”皮尔逊挺喜欢她，给她一笔奖学金让她继续学业，并且做他的研究生。

为 K·皮尔逊工作

在为 K·皮尔逊工作期间，大卫做的主要事情是计算一些复杂和困难的多重积分问题，以及计算相关系数的分布。这项工作使她写出了她的第一本著作《相关系数表》（Tables of the Correlation Coefficient），这本书最终在 1938 年正式出版。在那些年里，刀子所有

的计算工作都是靠一架名为“布伦斯维加”（Brunsviga）的手摇式曲柄计算机完成的。“我估计我大概摇了那架计算机两百万次……我常常碰到机器卡住这种倒霉的事，在我学会使用长针（来解决机器卡住）之前……这个机器一卡住，你只好跑去告诉教授，于是他就会数落你一顿，非常令人懊恼。所以有很多次，机器一卡住我就悄悄溜回家，没告诉他。”虽然她很钦佩皮尔逊，而且在他晚年大半的时间都陪着他，但在 20 世纪 30 年代的早期，大卫还是相当怕皮尔逊的。

大卫也是个很大胆的女孩子，常骑着摩托车参加越野赛。

> 有一次我撞上了一堵 16 英尺的高墙，墙头上还有玻璃。我被抛向半空中，伤到了膝盖。有一天我在办公室，心情沮丧，此时正好威廉 · S · 戈塞特进来。他说，“你以后最后改玩钓鱼吧。”因为他自己是个钓鱼的高手。他邀请我到他家中。在他亨敦（Henden）的家中有他、他的太太和几个孩子。他教我钓鱼，待我很亲切。

当 J • 奈曼与埃贡 • 皮尔逊开始形容费歇尔的似然函数时，大卫也在该大学学院，老皮尔逊认为埃贡研究的东西毫无意义，因此相当不悦。埃贡怕苦恼老爸，所以没有把他们第一份研究论文交给他父亲的期刊《生物统计》发表，反而与奈曼一起筹创另一份期刊《统计研究纪事》（Statistical Reserch Memoirs），共经营了两年（F • N • 大卫在上面发表了好几篇论文）。后来 K • 皮尔逊退休，埃贡接替他的父亲担任《生物统计》的主编，这时才把自己办的期刊停掉。

当这个“老家伙”（当时大家都这么称呼 K • 皮尔逊）被自己的儿子和费歇尔取代时，大卫当时也在。当年轻的 J • 奈曼刚开始做统计研究时，大卫就在那里。她回忆说，“我认为，20 世纪 20 年代至 1940 年间是统计学界生机勃发的时候，而我则从一个小人物的视角见识到了各路统计精英。”

大卫称 K • 皮尔逊是个绝对的演说家。“他讲得太棒了，你只能静静地坐在那儿，沉浸在他的演说中。”他对学生提问题打断他的讲话很耐心和宽容，即使有人指出他的错误也不要紧，他会很快纠正错误，然后继续讲下去。但另一方面，她觉得听费歇尔的演讲“是一件可怕的事，我什么都听不懂。我很想问他问题，但是当我真的提出问题时，他一看我是个女生就不屑回答我。”因此，她就坐在一个从美国来的男同学旁边，一有问题就推他的手臂说，“问他！问他！”“每次听完费歇尔的演讲，我总要上图书馆呆上三五个小时，想弄清楚到底费歇尔讲了些什么。”

1933 年，K • 皮尔逊退休，F • N • 大卫继续跟他做研究，成了他唯一的研究助理。大卫写道：

> K · 皮尔逊是个非同寻常的人。他已经 70 多岁了，但还整天工作，研究某些问题，有时候甚至会到早上 6 点才离开学校。有一次，当他正准备回家而我也正准备回家时，他对我说，“今晚你可以把椭圆积分的部分看一看，明天我们要用。”我当晚其实正准备和男朋友到切尔西（Chelsea）艺术厅参加舞会，但没有勇气告诉他。因此我还是和男朋友去跳舞，到了凌晨四五点才回到家，洗个澡之后就赶到学校去，看相关的资料做好准备，等皮尔逊 9 点左右到学校来。人年轻的时候总是好做傻事。

在 K • 皮尔逊去世前的几个月，F • N • 大卫回到了生物统计实验室与奈曼一起工作。当奈曼得知她还没有取得博士学位时，感到非常吃惊。在奈曼的催促之下，她把最后发表的 4 篇论文整理出来，提交出去当作博士论文。后来有人问她，在得到博士学位之后，你的地位有没有什么改变？她回答，“没有任何改变，我只是付了 20 英镑的入门费。”

回忆以往的那些日子时，她说，“我总以为他们让我加入，是为了使奈曼先生保持安静，但那段时间还是非常喧闹的。当时费歇尔在楼上，时常大声发表意见，奈曼在一边，而 K • 皮尔逊在另一边，此外戈塞特每隔一周也会来一次。”其实她对这些年的回忆过于谦虚了，而她自己也绝不是她所说的那种配角，“加入是为了让奈曼先生保持安静”。她所发表的统计学

论文，不论是在理论上还是在实践上，在很多领域中都大大提升了统计学的水平（其中有一篇更是非常重要，是她与奈曼联名写的，论述 20 世纪早期俄国数学家 A·A·马尔可夫（A. A. Markov）的某个定理的广义定理。在我的书架上，几乎每一学派统计理论的书上，都会把 F·N·大卫的论文当参考文献，可见她研究范围的广泛。

关于战争的研究

当第二次世界大战在 1939 年爆发时，大卫在国家安全部做研究工作，试图预测炸弹落在像伦敦这种人口中心时会有什么后果。预测的内容包括伤亡人数、炸弹对电力系统、饮水与污水管线系统的影响以及其它可能产生的问题，这些问题均可由她建立的统计模型估计出来。结果是，在 1940 年和 1941 年间英国对于德军向伦敦发动的闪电战，均做好了相关的准备，在及时抢救伤员的同时，还能维护主要的公共设施运转。

在战争快结束时，情况正如她所写的：

> 我坐着其中一架美国轰炸机，飞到安德鲁空军基地。我此行的主要目的是看看他们所造的第一批大型数码电脑……它就像个半圆筒形的活动式营房，长约 100 码，全部都是高架木板，你甚至可以在上面跑步。在两侧，大概每隔几英尺就有两个会眨眼的怪物，而天花板上除了保险丝什么都没有。每隔 30 秒左右维修人员就沿着木板巡视一遍，主要是抬头察看天花板上的保险丝……我回到英国以后，把看到的东西告诉一些人……他们则建议，“你最好是坐下来学习电脑编程语言。”我就说，“鬼才听你的！如果我这么做，我这辈子就只能做这个了，我不学，让别人去学吧！”

E·皮尔逊不像他爸爸那样喜欢权威式管理，他创造了一个新的惯例，就是系里的教授轮流当生物统计系的系主任。在轮到 F·N·大卫当系主任的时候，她正好开始写《组合机遇》（Combinatorial Chance），这本书后来成为一本经典的作品。该书详细地解释了复杂的计数法，也就是我们熟知的“组合数学”（combinatorics）。书中把原本极为复杂的观念，用简单通俗的方式陈述出来，从而使这些观念容易理解得多。当有人问起她这本书时，她回答：

> 在我一生当中老是陷入同样的困扰。我先是开始于一些事情，接着就会感到厌烦。我很早就有组合数学的想法，而且更早就开始这方面的研究，甚至在我认识巴顿（D. E. Barton，她的书的合作者，后来成为大学学院计算机科学系的教授）或给巴顿当老师之前……但我还是请他来跟我共同写这本书，因为我设想的事情也该了断了。因此我们一起写，他做了很好很深入的工作。他是个很不错的人，我们还一起写了许多论文。

她最后到了美国，成为加州大学伯克利分校的教授，还继奈曼之后，成为生物统计学系的系主任。1970 年，她离开伯克利到加州大学的河滨（Riverside）分校创办统计学系，并担任系主任。她在 1977 年 68 岁的时候“退休”，成为伯克利生物统计系很活跃的荣誉教授和研究人员。本章好多处引文出自于 1988 年对她的采访。她于 1995 年过世。

1962 年，F·N·大卫出版了一本书，书名为《赛局、上帝与赌博》（Games，Gods，and Gambling）。下面是她就为何写作该书所做的描述：

> 我年轻的时候学过希腊文……当时我有个从事考古研究的同事，当他一天到晚忙着在某个沙漠里东挖西掘的时候，我觉得自己也对考古学开始感兴趣了。不管怎么样，他曾对我说，“我在沙漠里走来走去，在地图上标示出可能有考古碎片的地方。凭这个地图我就知道应该在什么地方挖掘可能找到各种餐具的碎片。”考古学家对金银不感兴趣，只喜欢一些瓶瓶罐罐。我把他的地图拿来，仔细思考之后，发现这和我研究的德国 V 型轰炸机的问题很像。伦敦在这里，轰炸机的落点在另一个地方，而你想知道的是轰炸机的发射地点，这样你就可以假设一个双变量（bivariate）的正态平面，然后预测出几

个主轴。这就是我由碎片图得到的灵感。问题与问题之间似乎有某种共通性,很奇妙吧?
而且总共可以归纳成大约 6 种不同的类型。
弗洛伦斯·南丁格尔·大卫对所有这些类型的问题都有著述，做出过相当大的贡献。

第 16 章 非参数方法

在 20 世纪 40 年代，美国氰胺公司的化学家弗兰克·威尔科克森（Frank Wilcoxon）深为一个统计问题所困扰。针对不同化学处理的结果，他采用“学生”[18]t 检验和费歇尔的方差分析做假设检验，，进行比较。这是当时分析实验数据的标准方法，统计革命已经深入到了科学实验室，有关解释这些假设检验所用统计图表的书，已经摆到了每位科学家的书架上。但是威尔科克森所关心的，是这些方法常常表现为失效的情形。

他进行了一系列的实验，在他看来，这些实验中不同处理的结果显然是并不相同的。但是，有时候 t 检验显示了统计上的显著性，而有时候却没有。当进行一项化学实验时，常常碰到，在实验程序之初反应发生器（即化学反应进行的地方）并未充分预热，也会碰到某种特殊酶的反应力开始发生变化，结果使得实验结果似乎有误，常常是某个数据要么过大，要么过小。有时可以找到产生异常结果的原因，而有时虽然结果是一个异常值（outlier），显著地不同于其它结果，但又找不到明确的原因。

威尔科克森考察了 t 检验和方差分析的计算公式，意识到这些极端的异常值显著地影响了结果，导致“学生”t 检验统计量的数值比正常情形下的数值更小（一般而言，大的 t 检验统计量对应着小的 P 值）。这诱使他从观测值的集合中剔除异常值，用剩下的观测值计算 t 检验统计。这样一来，假设检验中的数学推导便出了问题。化学家如何才能知道一个数到底是不是异常值呢？必须剔除多少个异常值呢？当异常值被剔除之后，化学家还能继续使用那些基于标准检验统计量的概率图表吗？

弗兰克·威尔科克森着手搜集有关的文献，他确信那些发明统计方法的伟大数学家们早已注意到了这一问题。然而，他没有发现相关的参考文献。威尔科克森认为他找到了一个解决该问题的思路，但这一思路计算非常繁琐，要用到观测数据的组合与排列（前一章已经提到了 F·N·大卫的组合数学）。于是，他便着手寻找计算那些组合数的方法。

唉，这实在是太荒唐了！为什么要由一个像威尔科克森那样的化学家去研究这些简单而繁琐的计算方法呢？统计学界早应有人完成了这一工作!他于是又回到统计学文献中去找以前的论文，但他还是没有找到这种论文。他便寄了一篇论文给《生物统计学》(Biometrics）杂志（不要与 K·皮尔逊的《生物统计》(Biometrika）混淆），主要是想验证一下自己的数学方法。他并没有想过自己的研究会是一个原创性的工作，还想着审稿人一定知道文中内容早已在哪儿发表过了，从而拒绝他的论文，这样一来，也就等于审稿人告诉了他所需要的那些参考资料。然而，就审稿人和编辑们所知，这是一个原创性的研究，以前没有人思考过这一问题，他的论文在 1945 年发表了。

威尔科克森和《生物统计学》的编辑们都不知道，一个名叫亨利·B·曼（Henry B. Mann）的经济学家和俄亥俄州立大学（Ohio State University）一个名叫 D·兰塞姆·惠特尼（D. Ransom Whitney）的统计学研究生都在研究一个相关的问题。他们正试图给统计分布排序，这样一来人们便可以认为，在某种意义上，1940 年的工资分布“小于”1944 年的工资分布。他们找到了一种排序方法，但要用到一系列简单而繁琐的计数方法。

这促使曼和惠特尼设计了一个检验统计量，该统计量的分布也能用组合数学计算出来，与威尔科克森的计算类型一样。他们在 1947 年发表了一篇论文，介绍这种新方法，这已经比威尔科克森发表的论文晚了两年。很快便发现，威尔科克森检验（Wilcoxon test）和曼－惠特尼检验（Mann-Whitney test）密切相关，产生同样大小的 P 值。但是，这两个检验统计量引出了一些新的东西。直到威尔科克森发表之时，统计学界普遍认为，所有检验统计

[18] 实际上，运用施蒂格勒的误称定律进一步考证，威尔科克森并不是第一个提出非参数检验方法的人。K·皮尔逊在 1914 年的著作中似乎就已经提到了这种思想。但是，直到威尔科克森推出自己在这方面的研究，人们才充分地意识到非参数方法是统计学上的一次巨大革命。

量都是建立在数据分布的参数估计基础上的。但是新的方法是一种无需估计任何参数的检验方法，仅需要将观测数据的散点图与纯随机分布所预期的情形进行比较，这属于一种非参数检验（nonparametric test）[19]。

由此，统计学在 K·皮尔逊一些初步的想法之上迈出了革命性的一步，现在无需使用参数就可以处理数据分布的问题了。在西方，多数人都不知道，其实在 20 世纪 30 年代后期，苏联的安德烈·柯尔莫哥洛夫和他的一个学生 N·V·斯米尔诺夫（N. V. Smirnov）就发展出了一种不同的无需使用参数的分布比较方法。威尔科克森、曼和惠特尼的研究发展了数学研究的一个新领域，将注意力引致了有序秩（ordered ranks）的根本性质上，斯米尔诺夫－柯尔莫哥洛夫的研究成果也很快被纳入其中了。

进一步的发展

一旦在数学研究中出现了一个新的领域，就会有人用不同的方法去思考。在威尔科克森最初的研究后，很快就涌出了许多不同的替代方法。赫尔曼·谢诺弗（Herman Chemoff）和 I·理查德·萨维奇（I. Richard Savage）发现，威尔科克森检验可以看作是次序统计量（ordered statistics）的期望均值，他们还能将非参数检验扩展为关于不同基础分布（different underlying distribution）的一系列检验，都不需要进行参数估计。到了 20 世纪 60 年代早期，这类检验（现在被称为“非参数检验”（distribution-free tests）成了最热门的研究课题。一些博士研究生选择该理论中的某些小问题来做学位论文，一些会议专门讨论这种新的理论。威尔科克森也继续进行该领域的研究，提出了组合计算的更为精巧的算法，扩展了检验的应用范围。

1971 年，捷克斯洛伐克的雅罗斯拉夫·哈耶克（Jaroslav Hájek）写了一本权威的教科书，书中提出了该领域的一般性理论。他针对所有的非参数检验作了根本性的一般化，将一般化的方法与中心极限定理（the central limit theorem）的林德伯格－利维条件（Lindeberg-Lévy conditions）联系起来了。这正是数学研究中常用的方法。从某种意义上说，所有的数学实际上都是相互联系的，但是这些联系的准确性质和用于挖掘这些联系的见识，常常需要很漫长的时间才能显现。哈耶克于 1974 年去世，年仅 48 岁。

当弗兰克·威尔科克森试图将其在统计上的研究成果推广应用时，他放弃了最初的化学领域，而是在美国氰氨公司及其勒德勒实验分室（Lederle Labs divison）建立了一个统计服务小组。1960 年，他来到了弗罗里达州立大学（Florida State University）的统计系，成为一名倍受尊敬的老师和研究人员，指导了几名博士研究生。当他在 1965 年去世后，身后的学生和统计创新方法，仍然对统计学产生着重大的影响。

尚未解决的问题

非参数检验的发展促使人们在这一新领域进行了大量的研究。然而，在以前所用的参数方法与非参数方法之间，好像并没有什么明显的联系，因而还有两个问题尚未解决：

1. 若数据具有一个已知的参数分布，如正态分布，这种情况下我们采用非参数分析方法会有多不好？

2. 若数据不太适合采用参数模型（parametric model），那么数据必须偏离参数模型多远时，使用非参数方法才会更优？

---

[19] 实际上，运用施蒂格勒的误称定律进一步考证，威尔科克森并不是第一个提出非参数检验方法的人。K·皮尔逊在 1914 年的著作中似乎就已经提到了这种思想。但是，直到威尔科克森推出自己在这方面的研究，人们才充分地意识到非参数方法是统计学上的一次巨大革命。

1948 年，《数理统计学年报》的编辑收到了一篇来自塔斯马尼亚大学（the University of Tasmania）的一位不出名的数学教授的论文，这所学校位于澳大利亚南部的海滨小岛上。这篇杰出的论文一举解决了上述两大难题。那时，埃得温•詹姆斯•乔治•皮特曼(Edwin James George Pitman）已经在《皇家统计学期刊》上发表了 3 篇早期的论文，在《剑桥哲学学会会刊》（the Proceedings of Cambridge Philosophical Society）上发表了一篇论文，回过头去看，后一篇论文奠定了他后续研究的基础，但是它被人们忽略或是遗忘了。除了那 4 篇论文，在向《数据统计学年报》投稿时，已经 52 岁的皮特曼没有发表过其它的著作，也没什么名气。

E•J•G•皮特曼于 1897 年生于澳大利亚的墨尔本。他考入墨尔本大学（the University of Melbourne）念本科后，由于第一读世界大战而中断了学业，服了两年兵役后，他回到学校念完了本科。“那时，”他后来写道：“澳大利亚的大学没有数学方面的研究生院。”一些大学为优秀学生提供奖学金，到英国继续上研究生，但是墨尔本大学没有。“当我学习 4 年后离开墨尔本大学时，我尚未接受过研究方面的训练，但是我想我已经学会该怎么去学习和使用数学，可以去就应付所碰到的任何问题……”然而，首要的问题是要赚钱来养活自己。

塔斯马尼亚大学正要找人教数学，皮特曼去应聘而成为了一名数学教授。整个系就两个人，一位新来的教授和一位兼职计量。该系要为所有其它系的本科生上数学课，因此新教授忙着讲课，占去了几乎所有的时间。当理事会决定招聘一位全职的数学教授时，一位理事曾听说过数学有一悠闲的的分支叫做统计学，因此问应聘者是否准备讲统计学的课程（不管统计学到底是什么东西）。

皮特曼回答：“我并不能说我具备统计学的专业理论知识，但是如果被聘用，我将稍做准备，在 1927 年开出这门课。”他不具备统计的专业知识，也不具备统计理论的任何其它相关知识。在墨尔本大学，他学地一门高级逻辑学的课程，老师用了几次课来介绍统计学。正如皮特曼所指出的，“当时，也就是在那里，我认定统计学并不是我所感兴趣的东西，也永远不会为它而苦恼。”

年轻的 E•J•G•皮特曼在 1926 年秋天来到了塔斯马尼亚州的霍巴特（Hobart），只不过是一个本科生而已，却顶着教授的头衔。这是一个偏远的省级学校，根本感受不到身处伦敦和剑桥那种学术圈内的骚动。他写道，“直到 1936 年我没有发表过任何东西。之所以迟迟没有东西发表，主要有两个原因：一个是工作负担繁重，另一个是我所受教育背景的限制。”他的意思是说，他在数学研究方法上的训练不够。

到了 1948 年，当他将那篇非凡的论文投到《数理统计学年报》的时候，塔斯马尼亚大学数学系队伍有所壮大，已有一位教授（皮特曼）、一位副教授、、两位计量和两名助教。他们所开的数学课名目众多，既有应用数学方面的，也有理论数学方面的。皮特曼每周上 12 次课，周六也上课，同时获得了一些研究资助。从 1936 年开始，联邦政府为了促进澳大利亚高校的科学研究，每年拨出 30000 英镑进行资助。这些经费按人口在各州分配，因为塔斯马尼亚是一个较小的州，因此全校每年总共能得到 2400 英镑的资助。至于皮特曼能分到多少，他没有说。

慢慢地，皮特曼开展了多方面的研究，他发表的第一篇论文是关于流体力学中的一个问题。随后的 3 篇论文研究假设检验理论中几个特别的问题，这些论文本身倒并不怎么值得称道，但却是皮特曼的习作，探讨如何来发展自己的观点，怎样将数学的不同分支想到联系起来。

直到他开始撰写 1948 年那篇论文，皮特曼才建立起有关统计假设检验的性质以及过去的检验（参数方法）与新的检验（非参数方面）之间相互关系的一个清晰的逻辑框架。凭借着新方法，他解决了上述两大难题。

他的发现令人惊讶，甚至当原来的假设为真时，非参数检验也几乎与参数检验一样的棒。

皮特曼成功地回答了第一个问题：当我们知道参数模型和本应使用特定的参数检验时，如果还使用非参数检验，结果会有多差呢？皮特曼的答案时，根本不差。

第二个问题的答案更让人吃惊。如果数据不适合用参数模型，得差多远时使用非参数检验才会更好呢？皮特曼的计算表明，只需稍稍偏离参数模型，则非参数检验将远远地胜过参数检验。看起来，曾经深信别人早已做出了这个简单发现的化学家弗兰克·威尔科克森，似乎也是在无意中碰到了统计学中一块真正的点金石（philosopher’s stone）。皮特曼的结论表明，所有的假设检验都应该是非参数方法的。K·皮尔逊发现了带参数的统计分布，这仅仅是第一步，现在，统计学家们在解决统计分布的问题时，无需再为参数而烦恼了。

数学这东西往往是玄而又玄。在那些看似简单的方法背后，威尔科克森、曼、惠特尼和皮特曼对数据的分布作了一系列的假设，要理解这些假设或许又得花上一个 25 年的时间。第一个烦人的问题是由芝加哥大学（the University of Chicago）的 R·R·巴哈杜尔（R. R. Bahadur）和 L·J·萨维奇（L. J. （“Jimmie”） Savage）在 1956 年提出来的。几年前，当我将巴哈杜尔和萨维奇的论文给我的一位来自印度的朋友看时，他拿他们两人的名字匹配当戏谑，“Bahadur”一词在印度语是“勇士”（warrior）的意思，率先质疑非参数统计检验理论的是一名勇士和一个野蛮人（savage）。

巴哈杜尔和萨维奇所提出的那些问题实际上也正是源于异常值的问题，威尔科克森正是由该问题而首次提出了非参数检验方法。如果异常值极少，并且是完全“错误”的观测值，那么非参数方法将降低它们在统计分析中的影响。但是如果异常值系统性地污染了数据，采用非参数方法可能只会使分析更糟糕。我们将在第 23 章讨论有瑕疵数据分布（contaminated ditributions）的问题。

第 17 章 当部分优于总体时

在 K·皮尔逊看来，概率分布是可以通过收集有关数据来验证的。他认为，若收集足够多的数据，那么可以用来代表总体的相关数据。《生物统计》杂志的记者们从古墓中搜集到了数以百计的颅骨，灌入颗粒状物以测定颅腔的容量大小，然后将得到的几百个数据送给 K·皮尔逊。一名工作人员还深入中美洲的丛林中，测量了成百上千个当地土著居民的胳膊长度，这些数据也送到了 K·皮尔逊的生物统计实验室。

然而，K·皮尔逊所使用的方法存在一个根本性的缺陷。他获得的数据现在被称为“便利样本”（opportunity sample），都属于那些最容易得到的数据，并不能真正代表总体分布。他们测定的颅腔大小，都只是来自那些碰巧被他们发现而打开了墓穴，那些没有被发现的可能会与之大相径庭。

20 世纪 30 年代的早期，印度发现了一个便利抽样的典型案例。大包大包的黄麻堆到了孟买（Bombay）的码头上，准备装船运往英国。为了估计黄麻的价值，便从每包中抽取一些，黄麻的质量就由样本来确定。抽样是将一把中空的圆形刀片插入包中，再拔出来，刀片中央的空处便带出了少量的黄麻。在包装和上船过程中，外层的黄麻开始变质，而里面的被压得越来越紧，冬天的时候常常冻得结得一块。取样员将空心刀片插入包中时，由于中央更硬而发生偏离，所取的样品更多的是外层已经变质的黄麻。这种使得样本就会产生偏差，样本的质量偏低，实际上整包黄麻的质量要高出许多。

加尔各答市（Calcutta）总统学院物理系的普拉桑塔·钱德拉·马哈拉诺比斯（Prasanta Chandra Mahalanobis）教授经常引用这个例子（这是他在铁道公司工作时发现的，该公司将黄麻运往码头），说明为什么使得样本不可信。马哈拉诺比斯生于一个富裕的商人家庭，因此能够供他上本科和研究生，并且选择学习自己感兴趣的科学和数学。20 世纪 20 年代，他来到了英国，师从 K·皮尔逊和费歇尔。他的同学如 F·N·大卫只能靠奖学金生活，他却能一边上学，一边过着大地主般的生活。回国后，他担任了总统学院物理系的系主任。接着不久，他又在 1931 年用自己的钱，在自家的一处房产中建立了印度统计研究所（Indian Statistical Institute）。

在印度统计研究所，他培养出了一批卓越的统计学家和数学家，其中不少都在这一领域做出了重要的贡献，如 S·N·罗伊（S. N. Roy）、C·R·拉奥（C. R. Rao）、R·C·博斯（R. C. Bose）、P·K·森（R. K. Sen）和马丹·普里（Madan Puri），等等。马哈拉诺比斯的研究兴趣之一在于如何生成一个合适的、有代表性的样本数据。很明显，在许多情况下，几乎不可能得到一个总体的所有数据。例如，印度的人口是如此庞大，多少年来也没有人试图在一天之内搞一次全国性的普查，而这样的人口普查在美国曾经开展过。与此不同，印度的人口普查是在一年内完成的，全国不同地区分别在不同的月份开展。这样一来，印度的人口普查数据就不可能精确，在普查过程中会有出生和死亡、人口迁移，人口的自然状况也会发生变化。因此，没有人能确切地知道在特定的一天印度到底有多少人口[20]。

马哈拉诺比斯推断，如果能够收集到一个具有充分代表性的小样本，那么可以用它来估计总体的特征。在这一点上，我们有两种可能的方法：一是构造所谓的“判断样本”（judgment sample）。在判断样本中，所有关于总体的信息都被用来选择一个小的个体集合。这些个体分别代表总体的不同部分。有关多少人在看某一电视节目的尼尔森收视率排行榜（the Nielsen ratings），就是依据判断样本来排定的，尼尔森媒体研究所（Nielsen Media Research）根据社会经济状况和生活地区的差异，选择不同的家庭作为样本。

[20] 在美国，每 10 年进行一次人口普查，旨在登记全国在特定某一天的人口总数。然而，对 1970 年以来的人口普查的研究表明，这种全面登记仍将遗漏许多人，而另有一部分人却被重记。并且，这些被遗漏的人往往是出自某些特定社会经济背景的人群，因而也不能假定他们“类似”于已登记的居民。或许可以说，即使是美国，也没有人能够确切地知道在特定的一天到底有多少人口。

初看起来，判断样本似乎是获得大总体的代表性样本的好方法，但它有两个主要缺点。第一个是只有当我们确信对大总体具有充分的了解，可以将总体划分为能用一些个体来代表的几个子总体（specific subclasses）时，判断样本才具有代表性。既然我们希望通过样本来了解的问题，正是据以将大总体划分为几个匀质组（homogeneous groups）的依据，如果我们对大总体已经了解得这么清楚，可能就无需再进行抽样了。第二个问题更加麻烦，如果判断样本的估计结果是错的，我们无法知道该结果与真值到底相关多少。2000 年夏天，有人就批评尼尔森媒体研究所抽取的样本中西班牙裔家庭太少，因而低估了西班牙语电视的观众人数。

马哈拉诺比斯的解决办法是采用随机样本（random sample）。我们采用随机原则从大总体中抽取个体，由随机样本得到的数据很可能会错，但是我们可以用数理统计学的理论确定该如何最优地抽取样本并测定数值，以确保长期来看我们的数据将比其它数据更接近真值。并且，我们知道随机抽样概率分布的数学形式，可以计算总体那些待估参数的置信区间。

可见，随机样本要优于使得样本或者是判断样本，当然，这并不是因为它会保证得到正确的结果，而是因为我们可以计算一个数值区间，以较高的概率保证真值落入这一区间内。

新政与抽样

抽样的数学理论在 20 世纪 30 年代得到了迅速发展，其中一部分应归功于由马哈拉诺比斯领导下的印度统计研究所；一部分应归功于 20 世纪 30 年代后期奈曼发表的两篇论文；还有一部分应归功于一群年轻而富有朝气的大学毕业生，他们在美国实施新政的早期汇集于华盛顿。正是在这群在联邦政府商务部和劳工部任职的年轻人，热心于新政，提出了关于如何从总体中抽取样本的许多实际问题，并成功地解决了这些问题。

在 1932 年到 1939 年间，拿到学士学位的青年男女在跨出大学校门时，很难找到工作。这一切都是经济大萧条所造成的。在纽约州扬克斯市（Yonkers）长大的玛格丽特·马丁（Margaret Martin），毕业于巴纳德学院（Bamard College），后来成为美国预算局（the U. S. Bureau of the Budget）的一名官员，他写道：

> 当我在 1933 年 6 月毕业时，根据找不到工作……我的一个朋友比我晚一年毕业，找到了一个工作，在 B·奥特曼百货公司（B. Altman department store）当售货员，一周工作 48 小时，可赚 15 美元，对此她已经感到非常庆幸了。即使是那样的工作，也很难找到。巴纳德学院有一位负责就业指导的工作人员，也就是弗洛伦斯·多蒂（Florence Doty）小姐，我跑到她那儿咨询有关去一年叫凯瑟琳·吉布斯（Katherine Gibbs）的秘书学校受训的可能性，我不知道从哪儿能弄到这笔钱，但是我想在那儿学到一技之长来养活自己。多蒂小姐……是一个不太好相处的人，许多同学都对她敬而远之……她只回答了我几句，"我绝不造成你去学秘书课程，如果你去学了打字，并且让别人知道你会打字，那你以后就再也不能干别的了，只能打字……你应该去找一个专业性的职位。"

最后，马丁在奥尔巴尼（Albany）找到了第一份工作，成了纽约州失业安置局（the New York State Division of Placement and Unemployment）研究与统计办公室（the office of research and statistics）的一名初级经济师，这个工作成了她上研究生的一块跳板。

一些刚刚毕业的年轻 人直接进了华盛顿的政府机构。莫里斯·汉森（Morris Hanson）于 1933 年从怀俄明大学（the University of Wyoming）经济学本科毕业，去了普查局（the Census Bureau）。他凭着本科时学的一些数学和匆匆读过的奈曼的几篇论文，着手设计全国第一次失业普查。内森·曼特尔（Nathan Mantel）从纽约城市学院（City College of New York）生物专业毕业，去了国家癌症研究所（the National Cancer Institute）。杰尔姆·科恩菲尔德（Jerome Cornfield）毕业于纽约城市学院的历史专业，进入劳工部（the Department of Labor）

担任一名分析师。

那段时间，在政府工作倒是激动人心的，举国萧条，大部分正常的经济活动都停滞不前，可说是百废待举，华盛顿的新政府为此绞尽了脑汁。他们首先必须做的就是去了解整个国家到底已经糟糕到了什么程度，于是便着手对就业与经济活动开展各种调查。像这样试图准确地去判断这个国家到底是怎么了，在美国的历史上这还是第一次。很显然，这正是抽样调查发挥用武之地的时候。

这些干劲十足的年轻人首先要做的，就是说服那些不懂数学的人。劳工部在早些时候的一项调查显示，全国不到 10%的人口占有将近 40%的收入，这一结果受到了美国商会（the U. S. Chamber of Commerce）的公开指责，这怎么可能呢？调查人数还不到全国就业人口的 0.5%，而且这些人还是用随机方式获得的！于是，商会自己也进行了调查，主要是征求会员们对收入占有情况的看法。最后，劳工部调查的结果被商会认为是不准确的，拒绝接受，理由是那只不过是一堆随机的数据。

1937 年，政府想得到有关失业率的准确数据，同时国会授权在 1937 年进行失业普查。国会通过了议案，号召失业者填写登记卡，送到当地的邮局。那时，全国失业人口数估计在 300 万到 1500 万之间，仅有的较为可靠的数据是由纽约开展的几次随机调查所得到的。一群年轻的社会学家，在普查局的卡尔·戴德里克（Cal Dedrick）和弗雷德·斯蒂芬（Fred Stephan）带领下，认识到了可能会有许多失业者不填表，所得到的数据也可能包含着一些意想不到的错误。但他们还是决定，要在全国进行有史以来第一次严肃的随机调查。依据年轻的莫里斯·汉森对整个调查所作的规划设计，普查局从邮递线路中随机选取 2%，那些线路上的邮递员各自把调查问卷分发到所在线路的每一个家庭。

即使按 2%的比例抽样，普查局也被这样大量的调查问卷难住了。美国邮政服务局（the U. S. Postal Service）曾计划帮他们把问卷分类整理，并制作了一些原始的表格。问卷在最初设计时，还希望收集被调查人口统计和工作经历的详细资料，但是没有人知道该如何来处理这么大量的详细信息。别忘了，那时根本没有电脑，除了用“铅笔+纸张”绘制的表格之外，唯一可指望的就是手动的机械计算器。于是，汉森与耶日·奈曼取得了联系，当初他也是在奈曼的论文基础上完成了调查设计。用汉森的话说，奈曼指出，“我们不必知道或去探讨所有的细节，也不必弄清具体的关系如何”，只需为最重要的问题找到答案就行了。采纳了奈曼的建议，汉森和他的同事们抛弃了问卷中复杂而令人困惑的细节，只计算失业的人数。

在汉森的带领下，普查局作了一系列细致的分析，证实这种随机小样本调查的结论比起以前所用的判断样本要精确得多。最终，美国劳工统计局（the U. S. Bureau of Labor Statistics）和普查局都转入了以随机抽样为主要调查方式的新阶段。乔治·盖洛普（George Gallup）和路易斯·比恩（Louis Bean）又将这些方法引入了政治上的民意测验当中[21]。在 1940 年的普查当中，普查局还精心地设计了一些抽样调查计划。这时，普查局新来了一名年轻的统计学家，叫作威廉·赫维兹（William Hurwitz）。很快，汉森与赫维兹成了亲密的工作搭档和挚友，合作发表了一系列重要而有影响的论文，在 1953 年还合作出版了一本教科书《抽样调查方法和理论》（Sample Survey Method and Theory）（还有第三作者是威廉·马杜（William Madou））。汉森和赫维兹的论文与教科书在抽样调查领域里是如此的重要，并且极其频繁地被引用，以至于这一领域的许多人都认为有这么一个叫汉森·赫维兹（Hansen Hurwitz）的人。

[21] 在 20 世纪 60 年代末，我参加了一个研讨会，路易斯·比恩是演讲者之一。他介绍了最初的那些年，他和盖洛普开展民意测验，为政治候选人提供建议。后来，盖洛普面向公共领域，在多家报刊上发表专栏——盖洛普民意测验（Gallup Poll）。比恩继续做私人性质的民意调查，但他曾与盖洛普调侃，哪一天他也要建立一个专栏，取名为“急性子比恩民意测验（the Galloping Bean Poll）”。（译者注：gallop 与 Gallup 同音，意为飞驰的、急性的。）

杰尔姆·科恩菲尔德

新政期间，许多年轻人来到了华府，在政府机关和研究机构担任着重要的角色。不少人一直忙于发现新的数学与统计方法，根本顾不上去读研究生学位，最典型的要数杰尔姆•科恩菲尔德（Jerome Cornfield）了。科恩菲尔德在劳工统计局参与了最初的一些调查之后，来到了国家卫生研究所（the National Institutes of Health）。他和学界的领军人物合作发表了几篇论文，解决了个案控制研究（case-control studies）中的几个相关数学问题。他发表的科研论文内容广泛，涵盖了随机抽样理论、就业形态的经济学、鸡肉肿瘤问题、光合作用的问题以及环境毒素对人类健康的影响等诸多领域。他创立了许多统计方法，现在都已成为医学、毒物学、药理学和经济学等领域中统计分析的标准理论。

科恩菲尔德最重要的贡献之一，就是设计了弗拉明汉姆研究计划（the Framingham Study），并开展了初步的分析。这项始于 1948 年的计划最初是想以马萨诸塞州（Massachusetts）的弗拉明汉姆作为"典型小镇"（typical town），测定镇上每位居民有关健康状况的各种指标，然后对这些人进行多年的跟测。至今这项研究已经持续了 50 多年，期间曾发生过像"波林灾难"（Perils of Pauline）这样的事，因为政府减少预算开支，还常常试图降低对该计划的资助。但是现在这项研究已经成为分析饮食和生活方式对心脏病和癌症的长期影响的一份最主要资料。

为了分析弗拉明汉姆研究计划获得的头 5 年数据，科恩菲尔德碰到了几个基本问题，这些问题在以前的理论文献上还没有出现过。后来，他与普林斯顿大学的专业人员合作，一道把这些问题给解决了。其他人都沿着他所开创的理论发展方向继续写了不少论文，而科恩菲尔德为找到问题的解决办法而感到心满意足了。直到 1967 年，基于该项计划的第一篇医学论文发表了，他是其中的合作者之一。这篇论文研究了高胆固醇对得心脏病概率的影响问题。

1973 年，我和 J·科恩菲尔德同时参加一个会议，这是为国会某个专门委员会举办的系列听证会中的一场。某一天会议的间歇，有个电话找科恩菲尔德，原来是哥伦比亚大学的一位经济学家瓦西里·列昂惕夫（Wassily Leontief）打过来的，他说自己获得了诺贝尔经济学奖，并感谢科恩菲尔德在合作研究中所发挥的作用，正是他们的合作研究使他获得了这一奖项。他们的合作研究始于 20 世纪 40 年代后期，那时列昂惕夫曾跑去劳工统计局寻求帮助。

列昂惕夫认为，国民经济能划分为不同的部门，如农业、钢铁制造业、零售业，等等。每个部门都利用其它部门生产的原材料的服务，来生产某种原材料或服务，提供给其它部门，这种交叉关系能用数学中的矩阵形式来描述，常常被称为"投入——产出分析"（input-output analysis）。第二次世界大战后，当列昂惕夫刚刚对这一模型开始研究时，他曾到劳工统计局收集所需要的数据，劳工统计局指派了一名年轻的分析师协助他，这个人正是当时在那儿工作的杰尔姆·科恩菲尔德。

列昂惕夫可以将国民经济划分为几大部门，例如将所有制造业作为一个部门，也可以将各大部门进一步细分为若干个子部门。从数学原理上看，投入产出分析要求描述经济活动的矩阵必须存在唯一的逆矩阵，这意味着一旦获得了该矩阵，必须作为一个数学上"求逆矩阵"的去处。那时候，计算机并不普及，用手动式的计算器求逆矩阵非常的困难和繁琐。在我上研究生的时候，每个学生都必须练习求逆矩阵——我怀疑那简直是"净化灵魂"的一场仪式，记得当时求一个 5*5 阶矩阵，要花上好几天，大部分时间我是用来找错和改错。

列昂惕夫最初对经济部门的分类得到了一个 12*12 阶的矩阵，这样，杰尔姆·科恩菲尔德就要来求它的逆矩阵，看是否存在唯一的逆矩阵。大概花了他一周的时间，得到的结论是分类过粗，必须扩大经济部门的分类数目。于是，科恩菲尔德和列昂惕夫惴惴不安地对经济体系作进一步地细分，最后得到一个 24*24 阶矩阵，他们认为这是或许可行的最简单的矩

阵形式了。两人都知道，这一去处根本不可能由一个人完成。科恩菲尔德估计，计算一个 24*24 阶矩阵的逆矩阵，即使是一周工作 7 天，也要花上几百年的时间。

第二次世界大战期间，哈佛大学发明了第一台非常原始的计算机。这台计算机采用机械式继电器开关，还常常卡住。战争结束后，没有什么军事任务需要做了，哈佛大学正在找项目来使用这台怪物似的机器，于是科恩菲尔德和列昂惕夫决定将这个 24*24 阶矩阵拿过去，用这台叫作“马克Ⅰ号”（Mark Ⅰ）的机器来求它的逆矩阵，完成这一繁琐的计算。事后，当他们要为这一去处过程付费时，却被劳工统计局的会计部门制止。原来，那时政府部门有一项政策，货物可以购买，而服务不能购买。这一理论意味着，政府部门自身拥有各种各样的专业人员来为它服务，如果有什么事情要做，政府机构内部应该有能做这件事的人。

他们对政府中的那名会计解释说，理论上这件事有人能够做到，但是他活不了直到把这件事情做完那么长时间。那名会计对此非常同情，但文件就是那样规定的，但也无能为力。最后，科恩菲尔德想出了一个办法，顺利地解决了这个难题。方法是由劳工统计局开出一张购买固定资产的订单。什么固定资产呢？在发票上写的是劳工统计局从哈佛大学购买一个“逆矩阵”。

经济指数

在新政伊始进入政府机关的这群年轻人所做的工作，对整个国家来说仍旧极为重要。根据他们的研究成果建立起来的许多经济指标，现在已经成为对经济活动进行微调（finetune）时常用的参考指标。这些经济指标包括消费者价格指数（the Consumer Price Index，针对通货膨胀）、当期人口调查（the Current Population Survey，针对失业率）、制造业普查（the Census of Manufacturing）、普查局在 10 年一次的人口普查之间所作的中期调整（the intermediate adjustment），以及其它许多不那么出名的调查工作，所有这些都被世界各工业国所依仿效和沿期。

在印度新政府成立之初，P·C·马哈拉诺比斯成为首相贾瓦哈拉尔·尼赫鲁（Jawaharlal Nehrn）的一位私人朋友。尼赫鲁政府模仿苏联中央计划的做法，但是在马哈拉诺比斯的影响下，印度也经常开展一些深入的抽样调查，了解新国家真实的经济状况，以修正相关的经济政策。在苏联，各级官僚常常造出一些生产与经济活动的假数据，吹捧那些当政者，这又进一步造成了中央经济计划愚蠢地膨胀。在印度，总是可以得到对真实状况的准确估计，尼赫鲁和他的继任者们看了或许并不高兴，但也不得不慎重地处理。

1962 年，费歇尔来到了印度，此前受马哈拉诺比斯之邀他已经来过多次。但是这一次大为不同，全世界许多著名的统计大师都来到了印度，参加为印度统计研究所成立 30 周年而举办的庆祝盛会。费歇尔、奈曼、E·皮尔逊、汉森、科恩菲尔德以及其他来自美国和欧洲的众多嘉宾，云集印度。一系列的研讨会场面异常活跃，因为数理统计学还在蓬勃地向前发展，还存在着不少尚未解决的问题，同时，统计分析方法逐渐渗入到了各个科学领域之中，新的分析技术不断被提出，并接受检验。那时，全世界致力于统计学的科学学会已达 4 个，至少已有 8 种主要期刊（其中有一本是由马哈拉诺比斯创办的）。

会议结束后，嘉宾们各自回国。当他们回到家中，传来了噩耗——费歇尔在返回澳大利亚的途中，因突发心脏病在船上逝世，享年 72 岁。他的科学论文汇集成了 5 卷，所写的 7 本著作仍然对统计学的发展产生着影响，然而他卓著的原创性贡献却到此为止了。

第 18 章 吸烟会致癌吗？

1958 年，费歇尔在《百年回顾》（Centennial Review）中发表了一篇题为《香烟、癌症和统计》（Cigarettes, Cancer and Statistics）的论文，在《自然》（Nature）上发表了题为《肺癌与香烟？》（Lung Cancer and Cigarettes?）和《癌症与吸烟》（Cancer and Smoking）两篇论文。他后来把这几篇论文汇集在一起，编成了一个小册子《吸烟：关于癌症的争议及对有关证据的评论》（Smoking: the Cancer Controversy. Some Attempts to Assess the Evidence），还加上了一个内容广博的序言。在这几篇论文中，费歇尔（照片中的他常常是叼着一只烟斗）坚持认为，吸烟会导致肺癌的证据存在着严重的不足。

当时，不单是费歇尔在研究中批评了那些有关吸烟与癌症问题的研究，梅奥诊所（Mayo Clinic）的首席统计学家、美国生物学界泰斗之一的约瑟夫•伯克森（Joseph Berkson）也对这些研究的结论提出了质疑。耶日•奈曼也提出了反对意见，认为将肺部与吸烟联系起来的研究推理当中存在问题。费歇尔的批评最为强烈。在随后的几年中，由于证据渐多，伯克森和奈曼慢慢地也似乎认可二者之间的联系被证实了，费歇尔仍然强烈地反对，甚至指责一些主要的研究者篡改了数据，使许多统计学家都感到很尴尬。那里，烟草公司认为这类研究并不能说明什么问题，指出这只不过是一种"统计相关"，并不能证明吸烟会导致肺癌。从表面上看，费歇尔似乎同意他们这一观点，费歇尔的争辩火药味很浓，例如，下面是他一篇论文中的一段话：

> 一年前，《英国医学会期刊》（the British Medical Association's Journal）中登了一篇评论，得出了一个让人吃惊的结论：有必要运用当代所有的宣传手段让全世界人都详尽地了解吸烟的严重危害，这让我觉得有必要对此（那些试图证明吸烟与癌症之间关系的研究）作一个详细的分析。读这篇文章的时候，我觉得我很不喜欢"当代所有的宣传手段"这个词，而且在我看来，在这问题上应该有一个道德上的界限……为了让全世界一亿个烟民心存恐惧，而且还没说清到底该对这种舆论所反对的陋习担心些什么。，却花纳税人的钱而动用上了当代所有的宣传手段，对一个好公民来说，这实在是有点小题大做……

遗憾的是，在表示对使用政府宣传工具来传播这种恐惧的不满时，费歇尔并没有说清楚自己反对的到底是什么。这似乎印证了大家对他的看法，他就好像是那个反复无常的老头，只不过是不甘心扔掉自己那只心爱的烟斗罢了。1959 年，杰尔姆•科恩菲尔德与 5 位来自国家癌症研究所（the National Cancer Institute,NCI）、美国癌症学会（the American Cancer Society）和斯隆－凯特林研究所（the Sloan-Kettering Institute）的顶尖癌症专家一道，对所有已公开发表的研究作了一个回顾，撰写了一篇 30 页的论文。他们审查了费歇尔、伯克森和奈曼提出的反对意见，同时也探讨了烟草研究所（the Tobacco Institute，代表烟草公司的利益）的反对意见。他们由这场争论引申出阵一些更细致的推论，并且指出，有关证据压倒性地支持"吸烟是人类肺部表皮癌发生率迅速上升的原因之一"。

这篇论文平息了医学界关于这一问题的论争。尽管烟草研究所仍继续花钱在流行杂志上登整版的广告，质疑吸烟与肺癌之间的关系，认为它们仅仅是一种统计上的相关，但是在 1960 年以后，任何一本有名词的学术刊物上都不再有对这一发现提出质疑的文章了。该文发表之后不到 4 年，费歇尔便去世了，无法继续进行论战，也没有别人再掀起争议。

存在因果关系吗？

费歇尔的反对，难道仅仅是一个想安安静静地吸烟斗的老头在无理取闹呢，还是有着一定的道理？我读过他有关吸烟和癌症的论文，还将它们与他以前写的有关归纳推理（inductive reasoning）的性质、以及统计模型与科学结论之间关系的论文作了比较，发现了

一条前后一致的理论脉络。费歇尔所研究的是一个艰深的哲学问题——一个由英国哲学家伯特兰·罗素（Bertrand Russell）在 20 世纪 30 年代早期就提出来了的问题，这一问题抓住了科学思想的内核，但对许多人来说也许这并不算什么问题，即究竟何为“因果关系”？这一问题的答案绝对不那么简单。

许多读者也许都记得，那个满头白发、慈父般模样的罗素是一位世界著名的哲学家，在 60 年代，曾经公开批评美国政府介入越战。在那之前，他就被许多官员和学者认为是 20 世纪伟大的哲学家之一。他的第一部主要著作，是与艾尔弗雷德·诺思·怀特海（Alfred North Whitehead，比他早入道好些年）合写的，探讨了算术与数学的哲学基础问题，书名叫《数学原理》（Principia Mathematica）。这本书试图将数学的一些基本思想，如数字与加法，建立在集合论（set theory）所用的一些简单公理上。

罗素和怀特稍顷在这本书中运用了一个基本工具，就是符号逻辑（symbolic logic），这是一种新的研究方法，是 20 世纪早期的一项重大创造。读者可以回忆一下学过后亚里士多德逻辑（Aristotelian logic），例如，“人都难免一死，因为苏格拉底（Socrates）是人，所以他也将难免一死。”

尽管人类对亚里士多德式逻辑规则的研究已经大约 2500 年了，但相对而言这是一种没有什么用的工具。它过分强调那些很明显的事实陈述，建立一些武断的逻辑规则来判断什么符合逻辑，什么不符合逻辑，却未能模仿逻辑在数学推理中的运用，这恰恰曾是人们运用逻辑创造了新知识的一个领域。当学生们还在机械地背着像“苏格拉底也会死”、“乌鸦的羽毛是黑色”之类的逻辑分类规则时，数学家们正通过运用亚里士多德逻辑范畴之外的一些逻辑方法，发现着新的思想领域，如微积分。

在 19 世纪末和 20 世纪初，随着集合论和符号逻辑的发现，一切都发生了改变。从罗素和怀特海所用的最早形式来看，符号逻辑始于一些被称为“命题”（propositions）的思想元素。每个命题都有一个称为“T”或“F”[22]的真值，它们还能与一些代表“和”（and）、“或”（or）、“非”（not），以及“等于”（equals）的符号相结合与对照。因为每个原子命题（atomic propositions）都有一个真值，它们的任一组合也有一个真值，这个值可以通过一系列代数步骤来计算。在这个简单的基础之上，罗素、怀特海和其他人能够建立许多符号的组合，用来描述数字和算术，似乎还能描述各类的推理过程。

在所有的推理过程中，只有一种例外！人们似乎还无法创造出一套符号，用以表示“A 引致 B”（A causes B）。原因和结果的概念躲过了逻辑学家所作的各努力，总是无法套进符号逻辑的规则之中。当然，我们都知道“因果关系”意味着什么，如果我将一个玻璃杯摔到浴室地板上，那这一举动将使玻璃杯破碎；如果每当狗走错方向时主人就把它拉住，那这一举动将使狗学会走正确方向；如果农场主给庄稼施肥，那这一举动将使作物生长；如果一名妇女在怀孕前 3 个月服用催眠药（thalidomide），那这一举动将导致所生婴儿手足萎缩；如果另有一名妇女骨盆发炎，那是因为她在子宫里放了避孕器（IUD）[23]；如果某公司的高级

[22] 注意这类符号的抽象意义，“T”代表的当然是“True”（即“对”、“是”），“F”代表的是“False”（即“错”、“否”）。通过使用一些表面上没有意义的符号，数学家们能够考虑各种思想上的变化。例如，假设我们有三个真值“T”、“F”和“M”（代表“maybe”，即“或许”），这对数学意味着什么呢？通过采用这些纯粹的抽象符号，在符号逻辑中会导出各种引人入胜的复杂性，这一主题在过去的 90 年来一直是数学研究中一个非常活跃的领域。

[23]在 20 世纪 80 年代，美国联邦法庭受理了马德 V·G·D·瑟尔公司诉讼案（the case of Marder V. G. D. Searle），原告声称自己患病是源自她所用的子宫避孕器。原告提出的诉讼证据是，流行病学的证据表明，使用子宫避孕器的妇女得骨盆炎的概率明显增大。被告提供了一个统计分析，计算出了相对风险（即使用子宫避孕器妇女得骨盆炎的概率与未使用子宫避孕器妇女得骨盆炎的概率相除的比值）的 95%置信区间为（0.6，3.0）。陪审团不知道该怎么判了，最后法官裁定被告胜诉，并指出：“确信这一点特别重要：对于原因的推定，至少应基于它构成事件原因的一个合理的概率。”这项判决隐含着一个假定，即概率可以定义为个体事件发生的概率。尽管这一观点试图对“因果关系”和“统计相关”作出区分，但二者仍然是含糊的，而且这种含糊同样也发生在一些级别更高的法庭裁决当中，但是这种含糊，反映了因果概念中存在着的根本冲突，这

主管职位中女性极少，那是因为这家公司部分经理人员存在着性别歧视；如果我表兄弟脾气异常火爆，那是因为他属于狮子座的。

正如罗素在 20 世纪 30 年代早期所明确指出的，通常意义上的因果关系是一种相互矛盾的观念。不同的因果关系实例不能套用相同的推理程序，实际上，根本不存在所谓的因果关系，这只是一种流行的妄想，一个含糊的想法，它经不起纯粹理性（pure reason）的攻击。因果关系包含了一套互相矛盾的观念，在科学论述中几乎或根本没有价值。

实质蕴涵

为了取代因果概念，罗素从符号逻辑出发提出了一个清楚定义的概念，称为“实质蕴涵”（material implication）。通过使用原子命题的基本观念和一些联结符号如“和”、“或”、“非”与“等于”，我们就能产生“若命题 A，则命题 B”的观念，这与“非 B，则非 A”的命题是等价的。这听起来有点像贝叶斯定理中隐含的悖论（在第 13 章中我们介绍过），但还是有很大的差别，我们将在后面的一章中进行研究。

在 19 世纪后期，德国医师罗伯特·科赫（Robert Koch）提出了一组必要的假设，用来证明某种病原体（infective agent）将导致某种特定的疾病。这些必要假设是：

1. 只要病原体能够培养出来，疾病就会发生。
2. 只要疾病没有发生，则病原体一定没有培养出来。
3. 当病原体被消除，疾病就会消失。

虽然有点累赘，但是科赫给出了实质蕴涵的条件。在判断某种传染病是否由某种特定病菌引发时，这些条件可能是足够的。但是，对于吸烟和癌症之类的问题，科赫的必要假设就没有意义了。让我们来看看，肺癌和吸烟之间的联系在多大程度上符合科赫假设呢（从而检验了罗素的实质蕴涵是否适用）？病原体是吸烟史，疾病是肺部表皮癌。一些吸烟者并没有得肺癌，不满足科赫的第一个假设。一些得肺癌的人却声称他们没有吸过烟，若我们信其所言，则不满足科赫的第二个假设。如果我们将癌症类型限定为小细胞癌（small oat-cell carcinoma），那么不吸烟却得肺癌的人数几乎为零，因而也许满足第二个假设。如果我们拿掉病原体，也就是让病人停止吸烟，他还是可能得病，因此不满足科赫的第三个假设。

如果我们应用科赫假设（从而也就是应用罗素的实质蕴涵），符合这些假设的，只有那些由血液或者其它体液培育出的特定病原体所引发的疾病。但是，对于心脏病、糖尿病、哮喘、关节炎或者其它形式的癌症，这些假设就不再适用。

科恩菲尔德的答案

让我们回过头来，看看科恩菲尔德与 5 位知名癌症专家在 1959 年发表的那篇论文，他们逐一介绍了有关吸烟与癌症关系问题所作的研究[24]。首先是理查德·多尔（Richard Doll）和 A·布拉德福德·希尔（A. Bradford Hill）的研究[25]，发表于 1952 年的《英国医学期刊》（the British Medical Journal）上。多尔和希尔对英国死于肺癌人数的急剧增加感到十分吃惊，

也正是罗素早在 50 年前就曾经讨论过的。

[24] 5 位合作者分别为：国家癌症研究所的威廉·亨塞尔（William Haenszel）、美国癌症学会的 E·卡特勒（E. Culter）、约翰·霍普金斯大学卫生与公共健康学院（the School of Hygiene and Public Health, John Hopkins University）的亚伯拉罕·利林费尔德（Abraham Lilienfeld）、国家癌症研究所（the NCI）的迈克尔·希姆金（Michael Shimkin）和斯隆－凯特林研究所的厄恩斯特·温德（Ernst Wynder）。然而，这篇论文是由科恩菲尔德提出来并组织完成的，特别是他撰写了开展仔细检验和拒绝费歇尔意见的那几部分。

[25] 实际上，尽管费歇尔特别针对希尔和多尔的研究作了猛烈的抨击，但两人在将费歇尔的方法扩展到医学研究方面贡献卓著。希尔几乎是独自一人说明英国医药界，只有依据费歇尔的实验设计原则开展的研究才能获得有用信息。理查德·多尔，后来成为牛津大学的皇家医学教授（Regius Professor of Medicine），是将现代临床研究转化为统计模型的代表性人物。

于是搜集了数百名肺癌患者，将他们与一些非肺癌患者比较，这些病人也是同时进入同一家医院的，并且在其它方面相似（相同的年龄、性别和社会经济状况）。结果发现，肺癌患者中的吸烟人数几乎是非肺癌患者（在这种研究中，常常称为“对照组”（controls））中吸烟人数的 10 倍。到 1958 年底，这类研究另外还有 5 项，分别以斯堪的纳维亚、美国、加拿大、法国和日本的病人为研究对象，都得到了相同的结果：肺癌患者中吸烟人数大大地高于对照组中的吸烟人数。

这类研究被称为“追溯性研究”（retrospective studies）。

从一种疾病开始着手，向后看与这种病相联的有什么先决条件。这种研究需要有对照组（未患此病的其他组病人），用以断定恰恰是这些先决条件与此病有关，而不是病人某些更一般的特征。对于这种研究，常有人批评对照组可能与所研究的病例之间不相匹配。一项著名的追溯性研究是加拿大开展的，有关人造甜味剂（artificial sweetener）是否为膀胱癌（bladder cancer）病因的研究。结果表明，人造甜味剂与膀胱癌之间似乎存在着某种关联，但是通过对数据的仔细分析之后，发现这些病例几乎都来自社会经济地位较低的阶层，而对照组几乎都来自社会经济地位较高阶层。这意味着研究组与对照组之间不具有可比性。20 世纪 90 年代初期，耶鲁大学医学院（the Yale Medical School）的阿尔万·范斯坦（Alvan Feinstein）和拉尔夫·霍维茨（Ralph Horvitz）对如何进行这类研究提出了一些非常严格的规则，以确保研究组和对照组相互匹配。如果我们将范斯坦－霍维茨 规则（Feinstein-Horvitz rules）应用到这些针对癌症和吸烟关系的追溯性病例对照（case-control）研究之上，那么所有这些研究都不符合规则。

另一种替代的研究方法是事前研究（prospective study）。在这类研究中，事先选定一群人，详细记录他们的吸烟史，再跟踪他们以观察会发生些什么事。到 1958 年，已单独地进行了三次事前研究，第一次研究（同样是希尔和多尔所作，他们开展了第一个追溯性研究）选取了 50 000 名英国医生。实际上，希尔和多尔这项研究中并未对研究对象跟踪很长时间，而只是通过面谈了解了这 50 000 名医生的健康习惯，包括他们的吸烟习惯，跟踪调查 5 年之后，其中很多人真地患上了肺癌。这一研究确实说明吸烟与癌症之间存在联系。他们依据吸烟量的大小将这些医生分成了不同组，结果表明，吸烟越多的医生得肺癌的概率越大。这就是所谓的剂量反应（dose response），是药理学中产生反应的关键证据。在美国，哈蒙德（Hammond）和堆恩（Horn）对 187 783 名男子进行了一次前事研究（发表于 1958 年），他们跟踪调查了 4 个月，也发现了剂量反应。

然而，事前研究还存在一些问题。如果研究是小范围的，结论也许只是针对某个特定群体而言的，不能将它推广到更广泛的人群当中。例如，早期大部分的事前研究都以男性为研究对象，因为当时女性肺癌病例过少，无法开展研究。事前研究的第二个问题是，为了让事件（肺癌）发生得足够多，允许作有意义的分析，研究持续的时间必须很长。解决这两个问题，都需要跟踪大量的人群。大量的研究对象保证了可以将研究结果适用于更为广泛的人群。如果短期内事件的发生率很低，但只要跟踪的人数足够多，短期内同样能得到足够多的事件用以分析。

希尔和多尔第二个研究之所以选择医生有两个原因：一是医生对自己吸烟习惯的回忆比较可靠；二是他们近观过专业的医学训练，因此这群人中发生的所有肺癌病例肯定都会被记录下来。但是，我们能将针对那些接受过良好教育的专业医生得到的研究结论，推广到学历不及高中的码头工人上吗？哈蒙德和霍恩以近 200 000 的男性为研究对象，希望样本更具有代表性——而这可能会使所获精确信息更少。说到这里，读者可能会想起，某些人批评 K·皮尔逊的样本数据，理由是说那是一种便利样本。这些不也是便利样本吗？

为了回应这种反对意见，H·F·多恩（H. F. Dorn）在 1958 年研究了三个大城市的死亡证明书，然后对死者家属进行访问调查。这一研究选择了所有的死亡者，所以不能说是便

利样本。结果再次压倒性地证实，吸烟和肺癌之间存在着关联。然而，还是可能有人提出争议，会说对死者家属的访问调查存在不足。因为直到进行这项研究的时候，大家普遍都知道了肺癌和吸烟之间的联系，这样的话，与因其它病死亡者的家属相比，那些因肺癌死亡者的家属可能会对死者生前是否吸烟记得更为清楚。

这也正是大多数流行病学研究的情形，任何一项研究都可能存在着某些不足之处。对于任意一项研究而言，批评者总可以假想出导致结论偏差的各种可能情形。科恩菲尔德和他的合作者们搜集了 1958 年前针对不同国家、不同总体所作的 30 项流行病学研究。正如他们所指出的，这么多项针对各种总体开展的研究压倒性地一致，都得到了相同的结论，因而具有较高的可信度。他们对各种异议一项一项地进行讨论，也考察也伯克森的反对意见，表明了该如何用其中的某些研究来回应这些批评。奈曼曾经指出，若抽烟者活得比不抽烟者长而肺癌又是一种老年病的话，最初的那些追溯性研究可能就存在偏差。为此，科恩菲尔德等人用由这些研究中的病人所生成的数据表明，对这些病人的这种描述并不准确。

他们从两个方面讨论了便利样本是否具有代表性的问题。一方面，他们表明了所涉及的病人总体范围，增加了结论对不同总体都成立的可能性。另一方面，他们还指出，这种因果关系可能是源自生物学的基本原理，与病人不同的社会经济状况和种族背景无关，并且回顾了毒物学的研究，证实了吸烟对实验室动物和组织培养存在着致癌效应

科恩菲尔德等人的这篇论文是流行病学研究中有关如何求证病理原因的经典例子。尽管任一单项研究都存在着一些不足，但是随着证据越积越多，一项一项的研究使得同 不念旧恶结论越来越有说服力。

吸烟与致癌 VS. 橙剂

与上述现象形成对照的是，越战的老兵们认为战争中曾用的橙剂（Agent Orange，一种除草剂——译者注）对他们的健康造成了影响，使他们在后来的生活中备受折磨。有关的研究认定导致他们身体损害的原因，正是这种除草剂（herbicide）中所含的污染物，几乎所有这类研究都只是针对这一小部分以不同方式接触到了这种除草剂的人开展的。但是针对其它人群开展的研究却并不支持上述发现。在 20 世纪 70 年代，意大利北部的一个化工厂发生的一次意外事故，致使许多人接触到了剂量更大的该种污染物，但并没有产生长期影响。针对新西兰草场工人的研究却表明，那些接触了除草剂的人患一种特殊生育缺陷的可能性增大，但是这些工人大多数都是毛利人，毛利人从基因上说就容易出现这种特殊的生育缺陷。

有关吸烟与橙剂研究的另一个不同之处在于，对于吸烟，人们认为会引起的是一种很明确的病（即肺部表层癌），而由橙剂引起的问题很多，包括神经系统和生殖系统的一些病症。这种情况与毒物学中的一般发现相悖，在毒物学上，一般认为特定的药剂会导致特定类型的病害。对于橙剂的研究，没有得到任何有关剂量反应的迹象，当然，也没有充足的数据能判断这些人到底接触到了多大的剂量。总的来看，这一研究的结果含混不清，就是伯克森、奈曼和费歇尔等人的反对意见也无人问津。

通过对流行病学研究的分析，根据罗素的高度确切性要求和“实质蕴涵”的思想，我们已经非常深入了。现在因果关系从对人群总体许多有缺陷的调查推出，这种关系仅仅是统计意义上的，分布参数的变化可能源于某些特定的原因。但是，一些更为明智的研究者，可以通过综合大量的存在一些不足之处的研究，去发现一些共同的内在线索。

论文发表上的偏差

会不会是这些研究都经过挑选呢？观察者所看到的文献会不会只是从实际所做研究中精心挑选出来的一部分呢？又会不会是那些下面的研究发表了而负面的研究就没有发表呢？别忘了，并非所有的研究都能够发表。有一些论文会因研究者没有能力或不愿意而未能

做完，有一些论文会因为不符合杂志的规范而被编辑拒绝。特别是对所讨论的问题存在着争议时，编辑们常常倾向于发表那些容易为科学界接受的论文，而拒绝一些观点不易为科学界接受的论文。

这正是费歇尔提出批评的问题之一。他声称希尔和多尔最初的研究被改造过了，他多年力图让作者公布支持其结论的详细数据。而他们仅仅发表了论文的概要，但费歇尔认为这些概要掩盖了数据中实际所存在的不一致性。他指出，在希尔和多尔的第一个研究中，作者问吸烟者吸烟时是否吸入，这样将数据分为“吸入者”和“不吸入者”两类时，不吸入者得肺癌的多，而吸入者得肺癌的反而少些，希尔和多尔声称这一结果可能是因为部分被调查者没有弄清楚问题的含义。费歇尔对此很不以为然，并问他们为什么不公开真实的研究结论，让人们知道，虽然吸烟对你是有害的，但是如果你非得吸，与其不吸进去，还不如吸进去呢。

让费歇尔反感的是，希尔和多尔针对医生开展跟踪研究时，竟然将这个问题扔到了一边。那么，会不会还有其它什么问题也是精心挑选的呢？费歇尔很想知道。然而，更令他感到震惊的是，政府竟然不惜以大量的权力和金钱来将恐惧植入民众心理当中，他认为这种做法无异于纳粹利用传媒来操控民意。

费歇尔的答案

费歇尔也受到了罗素因果关系论的影响，他认识到实质蕴涵还并不足以描述大多数的科学结论，并写文章深入地讨论归纳推理的属性问题。他认为，如果很好地遵循了实验设计的有关原则，那么就有可能在某些特定研究的基础上得出关于生命的一般性结论。他还指出，实验设计中按随机原则将治疗方法分配给受实验乾，这种方法为归纳推论（inductive inference）提供了坚实的逻辑与数理基础。

那时，流行病学者都采用费歇尔所提出的实验设计分析工具，如他的统计估计与显著性检验方法。他们将这些工具用于便利样本的分析，在这类样本中实验处理的分配并非由研究之外的某种随机机制来决定，而是依据这些研究本身的复杂部分来确定。他的思索是，某些人吸烟而其他人却不吸，假定这是某种遗传基因的缘故，并且进一步假定，正是这种相同的基因结构导致了肺癌的产生。众所周知，多数的肺癌患者都具有家族性的特征。他因此提出，吸烟与肺癌之所以存在联系，大概是因为二者都由同一种因素所引起，即相同的基因结构。为了证明自己的推测，他收集了许多双胞胎的数据，结果表明，这些双胞胎要么两人都吸烟，要么都不吸，有着很强的家族性倾向。于是，他向其他人提出了挑战，要他们证明肺癌并非受相似的遗传基因所影响。

这场论战，一方是脾气火爆的天才费歇尔，他将统计分布的整个理论构建在了一个坚实的数学基础之上，正在作最后的一场战斗。而论战的另外一方是 J • 科恩菲尔德，他所受的正规教育只不过是一个历史学的学士，有关统计学的知识完全靠自学而来，忙于建立新的重要统计理论而没顾得上去拿更高的学位。费歇尔指出，不通过随机化实验，根本无法证明任何东西。科恩菲尔德却认为，有些现象本身就无法设计那种随机化的实验，但是承受着相关证据的累积也能说明一些问题。现在，两人都已经去世了，但他们学术思想的继承者尚在。在法庭上，当原告们举证自己受到了不公平的待遇时，这种争论便会时时现出；在分辨人类活动对生物圈的不利影响时，这种争论同样会扮演重要的角色；无论什么时候，一旦碰到医学中事关生死的重大问题，这种争论也必定会浮现出来。因果关系并不是那么简单就能够证明出来的！

第 19 章　如果您需要最佳人选……

1913 年夏末，乔治·W·斯内德克（George W. Snedecor）从肯塔基大学（University of Kentucky）获得了数学博士学位。他听说爱阿华（University of Iowa）有个数学教师的空缺，就收拾简单行李，搭车前往应征。不幸的是，他对爱阿华州所处位置一无所知，结果到了爱阿华州立学院（Iowa State College）的所在地——埃姆斯（Ames），而非爱阿华大学的所在地——爱阿华市。爱阿华州立学院的人告诉他，该校没有招聘数学老师，但该校已录取的有些学生数学背景不太好，问他是否愿意来教代数。6 年以后，他说服学校的人，应该让他设立一门关于统计方法新思想的课程。就这样，当费歇尔农业试验的第一篇论文问世时，斯内德克正在一所农业学校，并努力跟踪这些统计思想。

虽然斯内德克学的是数学，没有学过概率论，但他在埃姆斯研究这些新发展，并建立一个统计实验室。后来，他设立了美国的第一个统计系。他研究了费歇尔的论文，接着又阅读了其他人的著作，如皮尔逊、戈塞特（“学生”）、F·Y·埃奇沃思（F. Y. Edgeworth）、耶茨、冯·米泽斯等。斯内德克在原创研究方面贡献虽然不多，却是个伟大的编者。20 世纪 30 年代，他编写了一本教科书，书名就是《统计方法》（Statistical Methods）。起初，只是油印版，1940 年正式出版，立刻成为统计界的优秀教科书。他改进了费歇尔的《研究工作者的统计方法》，加进了一些基本的数学推导过程，并把类似的统计思想放在一起，还加了一大堆计算表，使读者不费什么力气就可以算出 P 值和置信区间。20 世纪 70 年代，有一篇评论文章指出，在所有领域的科学论文中，斯内德克的《统计方法》被引用的次数最多。

斯内德克又是一名很有效率的管理人员。他常邀请统计研究领域中的重量级人物暑期访问爱阿华州立学院。20 世纪 30 年代的多数夏季，费歇尔总会过来住上几个星期，讲讲学或担任顾问。从此，埃姆斯的统计实验室与统计系，成为世界上最重要的统计学研究中心之一。第二次世界大战前到此访问的教授们都是该领域的杰出人物。

格特鲁德·考克斯（Gertrude Cox，1900－1978）就是在这一时期进入了爱阿华州立学院。她原来梦想当一名传教士，到偏远的国度拯救灵魂。高中毕业后的大约 7 年时间内，她都在卫理公会教堂（Methodist Chruch）做社会服务工作。为了达到当传教士的心愿，必须完成大学学业。在大学学习期间，斯内德克使她相信，统计学比传教更有趣。因此，毕业以后，她继续跟随斯内德克，在统计实验室做研究。1931 年，她获得爱阿华州立学院颁发的第一个统计学硕士学位，斯内德克又聘用她在统计系任教。此时，她开始对费歇尔的实验设计理论特别感兴趣，因此，在学校里首次开设了实验设计方面的课程。后来，斯内德克替她在加利福尼亚大学（University of California）找到了一个攻读心理学博士的机会，她又在那里学了两年多的时间，获得博士学位之后，回到埃姆斯，斯内德克让她负责统计实验室的工作。

与此同时，著名的统计学家们仍然不断地访问爱阿华州的埃姆斯。威廉·科克伦（William Cochran）曾经停留过一段时间，教了一段时间的书。他和考克斯一起讲授实验设计（这时侯，已经开设了好几门这方面的课程）。1950 年，两人合写了一本教科书《实验设计》（Experimental Designs），这本教科书与斯内德克的《统计方法》一样，不但向读者讲述了统计方法，还介绍了该方法的坚实数学基础。书上有一组很有用的表，可以让实验人员针对具体情况修正实验设计、分析实验结果。《科学论文引用索引》（Science Citation Index）每年都会公布各个科学期刊上的论文引用书单，该索引用小号字体印刷，分为 5 列，《实验设计》每年都上榜，至少占上整整一列。

女性对统计学的贡献

读者或许已经注意到，除了弗洛伦斯·南丁格尔·大卫之外，本书到目前为止介绍的所有统计学家都是男性。统计学发展的早期，该领域主要是男性的天下。虽然也有很多女性在统计领域工作，但她们大都从事一些统计分析所需的繁复计算，实际上可以叫做"计算员"。正因为需要大量的计算，工具又只是手摇式的计算机，所以，这类繁琐的工作常由妇女来承担。女性比男性温顺、有耐心，大家比较相信她们，会让她们来检查计算结果是否正确。在 K·皮尔逊带领的高尔顿生物统计实验室（Galton Biometrical Laboratory）里，最典型的情景就是，皮尔逊带上几位男士四处走动，检查计算机算出的结果，或互相讨论深奥的数学理念，而女士们正在进行计算工作。

随着 20 世纪的发展，情况发生了变化。特别是耶日·奈曼，他帮助并鼓励很多女性，指导她们的博士学位论文，或与她们共同发表论文，并在学术圈里为她们寻找合适的职位。到了 20 世纪 90 年份工，当我参加全国统计学会的会议时，发现与会者约有一半是女性。在美国统计学会、生物统计学会、皇家统计学会和数理统计研究院，女性都有很杰出的表现。不过，与男性相比还不完全平等。许多统计学期刊上发表的文章，约有 30%的作者是女性或有女性参与，而美国统计学会的荣誉会员当中，只有 13%是女性。不过，这种性别方面的差距正在改变。20 世纪末的最后几年，占人类半数的女性已表现出她们所具备的较强数学能力。

但是在 1940 年，当斯内德克在火车上巧遇北卡罗莱纳大学（University of North Carolina）校长弗兰克·格雷姆（Frank Graham）时，情况还不是这样。他们坐在一起，谈论了很多。格雷姆曾听说过有关统计革命的情况，斯内德克正好是这方面的专家，他讲述了统计模型在农业及化学研究中的种种进展。格雷姆惊讶地得知，全美国居然只有爱阿华州立学院有正规的统计系，萨姆·威尔克斯（Sam Wilks，见第 20 章）在普林斯顿大学发展了一个数理统计小组，但还附属在数学系。亨利·卡弗（Henry Carver）所在的密西根大学（University of Michigan），情况也差不多。[26]格雷姆就火车旅行会谈中所了解的内容考虑了很多。

几星期后，格雷姆与斯内德克联系，表示自己已说服其姊妹学校——北卡罗罗莱纳州立大学（North Carolina State University），时机已经成熟，应该像爱阿华州立学院一样成立一个统计实验室，再发展成统计系。格雷姆询问斯内德克，能否介绍一位男士主持该部门的工作，于是，斯内德克坐下来列出了 10 个人的名字，认为他们可能会胜任该工作。他把考克斯叫进来，请她看看这份名单，并发表一下看法。她看完之后，问了一句："您认为我怎么样？"

于是斯内德克在推荐信里加了几句话："这些是我想到的最适合此工作的 10 位男士，但如果您需要最佳人选，我会推荐考克斯。"

后来，考克斯证明了自己不但是杰出的实验科学家和优秀的教师，还是一位出色的管理者。她组建的师资队伍，既是有声望的统计学家，也是优秀的教师。她深受学生的尊敬与爱戴，也深深地影响着学生们。我第一次遇见她时，是在美国统计学会的一次会议上，坐在我对面的是一个身材娇小的年长女士。当她说话的时候，眼睛里散发出一股热情，好像能燃起大家讨论主题的兴趣。不管讨论的是理论问题，或是实际应用问题，她的评论机智又风趣，叫人心服口服。当时我不知道她已经身患白血病，将不久于人世。她去世之后，她的学生每年夏天都会在各统计学会的传统联合年会上聚会，为纪念她而举办路跑，并筹措以她名字命

[26] 亨利·卡弗（1890－1977）在数理统计的发展上，是个孤独的先驱，也是一个值得尊敬的学术导师。从 1921 到 1941 年，他在密西根大学指导了 10 位博士研究生，他给学生的论文题目全部与数理统计有关。1930 年，他创办了《数理统计年报》（Annuals of Mathematical Statistics），1938 年，协助成立了"数理统计研究院"（Institute of Mathematical Statistics），该研究院是赞助年报的学术机构。《数理统计年报》后来获得极高的评价，我们将在第 20 章谈到。

名的奖学金。

1946 年，由于考克斯的“应用统计系”非常成功，所以格雷姆终于能在建在教会山上的北卡罗莱纳大学设立一个数理统计学系，不久又成立了生物统计系。从此之后，北卡罗莱纳州立大学、北卡罗莱纳大学与杜克大学（Duke University）成为统计研究的“铁三角”，很多私人研究公司也都听从这几所学校专家的意见。考克斯创建的统计世界，使她的老师斯内德克的成就相形见绌。

经济指标的发展

在美国联邦政府的统计部门，妇女扮演了非常重要的角色，她们分别在普查局（the Census Bureau）、劳工统计局（the Bureau of Labor Statistics）、国家卫生统计中心（the National Center for Health Statistics）及管理预算局（the Bureau of Management and Budget）等部门身居要职。其中职位最高的是珍妮特·诺伍德（Janet Norwood）女士，她于 1991 年退休，当时是美国劳工统计局的局长。

诺伍德女士就读于道格拉斯学院（Douglass College），这是拉特格斯大学（Rutgers University）的女子分校，位于新泽西（New Jersey）新布朗斯维克（New Brunswick）。当时美国正式参加第二次世界大战，诺伍德的男友必须入伍从军，于是他们决定先结婚，当时诺伍德 19 岁，伯纳德·诺伍德（Bernard Norwood）先生 20 风。婚后诺伍德先生并没有立刻被征调到海外，因此俩人仍能见面。但是，这桩婚姻却对道格拉斯学院这样的封闭环境造成一些困扰。在此之前，校园里从来没有已婚的学生。对男性来访者的限制性规定要用在她先生的身上吗？她离开学校到纽约探望先生，必须通知学生家长吗？这些都是由诺伍德女士首开先例的。1949 年，她获得塔夫茨大学（Tufts University）的博士学位，成了该校有史以来最年轻的博士。她自己写到：“接二连三地，在我工作过的几个岗位上，我总是第一个被提拔的女性。”她是美国劳工统计局的第一位女局长，从 1979 年任职到 1991 年退休。

1979 年联邦政府任命她为局长的时候，可能对她了解的并不太多。在诺伍德女士上任局长之前，劳工部有一项惯例，就是派一位熟悉政策事务的代表，出席所有由劳工统计局召开的记者招待会。诺伍德上任后，通知部里的代表，以后不必出席这类记者招待会。她认为，局里拿出去的各种经济资料，不但内容上应该准确，而且应具有无党派性，连形式上都应该如此，她要求局里的所有活动，都尽可能避免行政干扰。她说过：

> 我发现把下面这件事情讲清楚很重要，那就是：在遇到重大问题时，要相信原则。……在政府部门做事，应该主张并坚持独立性。……不过，这并不容易做到。例如，当必须对总统的意见进行修正时，应该怎么办？这时，必须修订。

诺伍德女士与她丈夫都是经济学博士。她们结婚的头几年，尤其在她丈夫参与研究欧洲共同市场的相关制度时，她并没有外出工作，只是在家教养两个孩子，偶尔写一些学术上的文章，让自己保持活力。后来，全家定居于华盛顿，他们的小儿子也开始上小学，诺伍德女士就出来找事做。她想找的工作是要能有几个下午不必上班的那种，这样当孩子放学回家时能照顾他们。劳工统计局有这样的工作机会，每周有三个下午在家。

劳工统计局在劳工部里是个小单位，在政府部门中很少制造什么大新闻。与白宫和国务院的刺激性工作相比，这种小局里会有什么事？事实上，在整个政府机器里，它是个很重要的齿轮。政府的工作必须靠资讯，那些由各地赶到首都华盛顿参与新政（New Deal）的聪明青年男女，不久就发现，要建立适用的政策，必须充分掌握全国或各州的经济状况，但在当时，根本没有这类信息。新政的一项重要改革，就是设立必要的机构，来提出有关国家经济发展的重要资讯。劳工统计局一方面进行必要的调查工作，以生产该类信息，另一方面，则对其他部委，如普查局，收集的数据进行分析。诺伍德女士于 1963 年进入劳工统计局，

1970 年就得到晋升，负责消费者价格指数的编制。消费者价格指数有很多用途，可用来衡量社会保障的支付，追踪通货膨胀的现象，调整从联邦政府到各州政府的大部分转移支付。1978 年，在诺伍德女士亲自策划与监督下，劳工统计局将消费者价格指数做了一次重大的修订。

诺全德女士担任局长之后，劳工统计局统计的消费者价格指数及其他系列指数，都牵涉了一些较为复杂的数学模型和若干个相当难懂的参数，这些参数虽然在经济模型中具有意义，但对那些缺乏经济数学训练的人来说，却很难解释。

报纸在引述消费者价格指数（CPI）时，经常会有“上个月的通货膨胀率上升了百分之0.2”这类的说法。但是，消费者价格指数是一组很复杂的数字，反映的是全国不同地区和不同经济部门的价格形态变动。它从“市场篮子”（market basket）的概念开始，“市场篮子”指一个典型家庭可能购买的一组货物和服务。在组合出该组货物和服务之前，必须先经过抽样调查，看看一般家庭到底会买一些什么东西，以及多长时间买一次。计算时，对不同的货物和服务，要赋予不同的权数（weight），因为一个家庭每周都要买面包，但好几年才买一次汽车，至于买房子的次数就更少了。

“市场篮子”及其权数一经确定，劳工统计局就派出人员，用随机方式抽选商店，并在选中的商店中记录所列商品的现价。然后，他们再把记录下来的价格，依照加权方式计算出一个总数，在某种意义上说，这个数字就代表了给定规模家庭的月平均生活费。

从理论上看，用指数来描述某种经济活动的平均形态，是一种很容易理解的想法，但要构建这样一个指数，就不那么容易了。对于市场上的新产品（如家用电脑）应该如何计算？如果某种产品的价格过高，消费者转而选择其他类似产品（例如买酸奶酪而不买酸奶油）又该怎么处理？消费者价格指数和其他用来度量国家经济运行是否良好的指标都要定期检查。诺伍德女士亲自督导了上一次消费者价格指数的重要修订，以后还会有人再做同样的事。

消费者价格指数并不是衡量国家经济状况的惟一指标，还有其他指数用来描述生产活动、存货及就业形态。还有一此社会指标，如监狱罪犯的估计数等这些与非经济活动有关的参数。但实际上，这都是 K•皮尔逊意义上的参数，是概率分布或数学模型的一部分，它们描述的不是具体的可观测事件，但又决定着可观测事件的形态。因此，美国没有一个家庭，每月的生活费正好等于消费者价格指数。同样的，失业率也不能描述实际失业人口，因为这个数字每小时都在变化，例如，什么人属于“失业人口”？从未工作过、也不打算找工作的人算不算？休假 5 周、领着离职金、正准备从前一家公司跳槽到另一家公司的人算不算？若有人每周只打算工作几小时，算不算？经济模型的世界中，对这类问题总是给出武断的答案，所牵涉到的众多参数永远不能确切地观测到，但它们彼此作用、互相影响。

在推导经济指标与社会指标时，可没有像费歇尔这样的天才，能够建立起最佳的标准。在每一个个案中，我们都设法把人们之间的复杂影响简化成一小组数字。不得不做出武断的决定。美国进行第一次失业普查时，只对户主进行计算（大部分是男性），而现行的失业率调查，则包括前一个月内想找工作的所有人。在修订消费者价格指数时，对武断程度差不多的定义存在着不同意见，作为督导者，诺伍德女士必须在它们之间求得一致，但永远会有诚恳的批评者就这些定义提出反对意见。

理论统计界的女性

本章提到的考克斯与诺伍德，扮演的角色主要是老师与管理者。20 世纪后半叶，妇女对理论统计学的发展也起到了重要作用。第 6 章介绍过蒂皮特，他的第一条极值渐进线能用来预测“百年难得一见的洪水”。这种统计分布有个改良版，称为“威布尔分布”（Weibull distribution），在航天工业中有很重要的用途。但威布尔分布有个问题，它不满足费歇尔的

正则性条件，因此，没有一种最优方法来对参数进行估计。后来，北美罗克韦尔（Rockwell）公司的南希・曼布之间有某种关联，因而，发展出一套方法，目前应用于该领域。

威斯康辛大学（the University of Wisconsin）的格雷斯・沃赫拜（Grace Wahba）女士采用一组特殊的曲线拟合法，叫做“样条拟合”（spline fits），并发现了支持当今样条统计分析的理论公式。

20 世纪 60 年代末，部分统计学家与医学家组成了一个委员会，他们设法研究三氟溴氯乙烷（halothane）这种麻醉剂的广泛使用，是否是病人肝衰竭发病率增加的原因，伊冯娜・毕晓普（Yvonne Bishop）女士是该委员会的成员之一。由于大部分数据以记录事件次数的形式出现，因此分析结果令人困惑。在此之前的 10 年间，很多人试图像研究三氟溴氯乙烷那样，制作一种复杂的多维计数表，但都没有特别的成效。这些研究人员曾经建议，应该用类似费歇尔的方差分析法去建立这样的表，但这项工作并未完成。后来，毕晓普女士接手了这项研究，检验了一些理论上的分歧点，并建立起估计与解释的准则。她把三氟溴氯乙烷研究得到的方法加以修饰之后，出版了一本权威性的著作。这个方法后来被称为“对数线性模型”（log-linear model），如今成为大部分社会学研究中首先要做的一个标准步骤。

从斯内德克和考克斯那时开始，“最佳人选”经常是女士。

第 20 章 朴实的德克萨斯农家小伙

20 世纪 20 年代末，塞缪尔・S・威尔克斯（Samuel S. Wilks，1906－1964）离开德克萨斯州（Texas）的家庭农场，到爱阿华大学读书。当时数学研究的工作主要是提升抽象之美。一些纯抽象领域，如符号逻辑（symbolic logic）、集合论（set theory）、点集拓扑学（point set topology）与超穷数理论（the theory of transfinite numbers）等，流行于美国各大学。由于过于抽象，使得任何与实际问题有关的灵感，哪怕只要沾到一点点边，都被抛到九霄云外。不少数学家一头栽进古希腊数学家欧几里德（Euclid）声称作为数学基础的公理当中。他们发现，在这些公理背后，还存在一些未被说明的假设。于是，他们设法去除这些假设，直接探索逻辑思维的基本构成元素，使自己沉浸在引人注目、但看似自相矛盾的想法中。如“填满空间的曲线”、“一个同时到处接触且不接触的三维体”等等。他们研究无穷大的不同阶数（order of infinity）及分数维度的“空间”。数学处于纯抽象思想之波四处席卷的高潮，没有丝毫现实世界意味。

没有任何地方像美国大学的数学系那样远离现实，深入抽象。美国数学学会（the American Mathematical Society）发行的刊物，是公认的全世界最顶尖的数学期刊，美国数学家在抽象的世界里一直往前走。正如威尔克斯几年后感伤地表示，这些数学系就像希腊神话中的海上女妖，不停地把全国最优秀的研究生引诱过去。

威尔克斯在爱阿华大学上的第一门研究生数学课程是由 R・I・穆尔（R. I. Moore）讲授的，他是本校数学教授中最有名的。穆尔讲授的是点集拓扑学，这使威尔克斯接触到抽象世界之美。穆尔毫不讳言自己看不起应用领域，他坚持认为应用数学与洗碗、扫街处于同一水平。从古希腊开始，这种态度就流行于数学界了。传说有一次，欧几里德对一个贵族小孩子讲解某个定理的美妙证明，老师虽然满腔热情，学生似乎无动于衷，反而问这有什么用。欧几里德听了，叫来一个奴隶，吩咐说：“给他一个铜币，他好像一定要从所学的知识里得到好处。”

威尔克斯后来转向实际应用领域，是由于博士学位论文的指导教师埃弗里特・F・林奎斯特（Everett F. Linquist）。当时，威尔克斯正在进行博士论文的先是工作，而林奎斯特曾研究过保险数学，对新发展出来的数理统计学很有兴趣，因此，从中为威尔克斯推荐了一个问题。那里，大家对数理统计的评价并不高，至少在美国和欧洲各大学的数学系里如此。费歇尔先生的开创性大作多半发表在一些“非主流”期刊里，如《爱丁堡皇家学会哲学学报》（Philosophical Transactions of Royal Society of Edinburgh）。而《皇家统计学会期刊》与《生物统计》都刊登一大堆制成表格的统计数字，因而受人轻视。亨利・卡弗已经在密西根大学着手创办一份新期刊《数理统计年报》，但对大多数的数学家来说，其程度还是太低，不能引起他们的注意。林奎斯特发现，教育心理学用到的测量方法中，有个很有趣的数学问题，就建议威尔克斯试试。后来威尔克斯把问题解决了，并以此作为博士论文，最后发表在《教育心理学期刊》（Journal of Educational Psychology）发。

对纯数学领域的人来说，这件事算不上什么成就。教育心理学的东西，引不起他们多大兴趣。其实博士论文也只是踏入研究工作实验性质的第一步，很少期望学生在博士论文里就做出重大贡献。后来，威尔克斯到哥伦比亚大学做一年的博士后研究（以增加他处理重要抽象数学概念的能力）。1933 年秋季，他受聘到普林斯顿大学，担任数学讲师。

统计在普林斯顿

普林斯顿大学的数学系也和美国其他大学的数学系一样，沉浸于许多冷酷、优美的抽象数学概念中。1939 年，普林斯顿高级研究院（Institute of Advanced Studies）成立。高级研究

院的第一批研究员中有 H・M・韦德伯恩（H. M. Wedderburn），他致力于将所有的有限数学群（finite mathematical groups）完全一般化。研究院还有赫尔曼・韦尔，他以无维度向量空间（nondimensional vector space）的研究出名，库尔特・格德尔（Kurt Gödel）发展出元数学代数（algebra of metamathematics）。这些人的风格影响到普林斯顿大学的教师们，这些教授本来就是世界知名的数学家，其中最突出的是所罗门・列夫契兹（Solomon Lefshefz），他打开了通往代数拓扑学（algebraic topology）这一新的抽象领域之门[27]。

尽管整个普林斯顿大学系都偏向抽象数学，对威尔克斯来说，幸好系主任是卢瑟・艾森哈特（Luther Eisenhart），他对所有的数学领域都很感兴趣，并且鼓励年轻教师依照自己的爱好进行研究，艾森哈特聘用威尔克斯到普林斯顿大学，就是认为数理统计是有发展潜力的学问。威尔克斯带着太太来到普林斯顿，追寻应用数学的远景，这使得他和数学系的其他教师们相比与众不同。他是个温和的战士，他那德克萨斯农家小伙的质朴，可以让任何人解除武装。他感兴趣的是个性的人，也能说服别人听从他的观点。他又是一个相当优秀的组织者，能安排各种活动去完成难以达到的目标。

当别人还在设法了解某个问题时，威尔克斯通常已经能直接切入该问题的核心，并想出一些可能的解决方法了。他的工作态度非常认真，也能说服别人像他那样努力工作。抵达普林斯顿大学不久，他就成为《数理统计年报》（也就是卡弗创立的那份统计期刊）的主编。他建立了论文发表的标准，并带领研究生一起编辑这份期刊。有位新来的同事约翰・图基，本来对抽象数学比较感兴趣，但威尔克斯说服他加入到他所从事的统计研究中。威尔克斯带过的许多研究生，第二次世界大战后纷纷在其他大学成立统计系，或在统计系任教。

威尔克斯的博士论文，处理的是教育心理学中的问题，因此，他有机会参与教育测试服务（Educational Testing Service）工作，帮助制定出抽样程度和评分方法，用于大学入学和其他学校的考试。他建立的理论工作，使得不同加权结构的计分方法仍然可以得到类似的结果。他和贝尔电话实验室（Bell Telephone Laboratories）的沃尔特・休哈特[28]也有联系，休哈特正开始把费歇尔的实验设计理论用于工业产品的质量控制上。

统计与战时事务

20 世纪 40 年代，威尔克斯最主要的工作，可能是在华盛顿担任海军研究局（Office of Naval Research）的顾问。他认为，实验设计法可以改善武器的使用效果，刚好海军研究局的人容易接受他人的建议。在美国参加第二次世界大战时，陆军与海军准备将统计方法应用在美国式的作用研究当中。在国防研究委员会（National Defense Research Council）之下，威尔克斯建立了普林斯顿统计研究小组（Statistical Research Group-Princeton，简写为 SRG-P）。这个研究小组招聘了一批聪明的年轻数学家与统计学家，其中很多人战后仍对科学有重大贡献。该小组的成员包括：约翰・图基，他把整个研究重心都转到应用上；弗雷德里克・莫斯特勒（Frederick Mosteller），他在哈佛大学设立了几个与统计有关的院系；西奥多・W・安德森（Theodore W. Anderson），他写的多变量统计教科书，后来成为相关领域的圣经；亚力山大・穆德（Alexander Mood），后来在随机过程理论上有重大的进展；查尔斯・温莎（Charles Winsor），他整个估计方法领域享有盛名；等等。

---

[27] 级研究院还有一位是艾伯特・爱因斯坦（Albert Einstein），但是他是物理学家，他的技能比穆尔所说的“打扫街道”复杂一点儿，他的研究搀入了相当多“实际生活”的应用。

[28] 当今，产业界里几乎每一个质量控制部门，利用的都是休哈特图（Shewhart chart），来追踪产出的变动。然而，沃尔特・休哈特（Walter Shewwhart）这个名字，也是施蒂格勒（Stigler）误称定律的一个例证。休哈特图的实际数学公式，最初似乎是戈塞特（“学生”）先提出来的，甚至在乔治・U・尤尔（George Udny Yule）的早年教科书上也出现过，休哈特只是表明如何把它应用在质量控制上，并作为一个有效的方法论加以推广。

安德森在普林斯顿统计研究小组工作的时候，还是个研究生，他提到了当时为了找出一种毁坏地雷的方法而进行的种种尝试。就在进攻日本本土的日子越来越近的时候，美国陆军得知日本已经开发出一种非金属地雷，已知的探测工具无法测到它。日本人将在海岸线上，以随机形态在可能的入侵路线上布满这种地雷。仅这种地雷造成的死伤人数据估计将高达数十万，因此，亟需一种可以毁坏这种地雷的方法。在此之前，欧洲曾尝试过从飞机上丢炸弹来引爆地雷，但没有成功。安德森与研究小组的其他成员曾组织在一起，设计利用引爆绳索来毁坏这种地雷的试验。依据安德森的说法，实验数据计算的结果显示，这种方法不可能有效毁坏地雷，这也是导致美国在日本投下原子弹的原因之一。

该小组也研究一种用在防空火炮上的近爆引管（proximity fuses），近爆引管本身会发出雷达信号，当发现目标接近时会自动引爆。此外，他们还协助开发出第一个会自动飞向目标的精巧炸弹（smart bombs）、研究测距仪（range finders）和各种不同各类的炸弹。普林斯顿统计研究小组的成员，不断地为全国各地的军事设施或军工实验室设计实验、分析数据。后来，威尔克斯又在哥伦比亚大学，协助组建了第二个统计研究小组（Statistical Reaserch Group-Princeton，Junior，缩写为 SRG-Pjr），这个小组的成果之一就是“序贯分析”（sequential analysis），这是一种当实验还在进行时，就可以对实验设计进行修订的方法。序贯分析所允许的实验修正，涉及每一个被检验的处理步骤。就算是最审慎的实验设计，得到的结果有时也会显示出，原先的设计要做一些变动，以使实验结果更为完整。序贯分析的数学理论会使科学家知道，在不影响结论有效性的 情况下，什么样的修订可行，什么样的修订不可行。

序贯分析研究从一开始就被列为最高机密，直到战争结束若干年后，参加这项研究的统计学家都不能对外发表论文。20 世纪 50 年代，第一批有关序贯分析及其“近亲”——序贯估计（sequential estimation）的论文发表之后，激发了其他人的想象力，整个领域迅速发展起来。今天，统计分析里的序贯法（sequential method）已在工业产品质量控制、医疗研究、社会学研究等领域广为应用。序贯分析只是威尔克斯及其统计研究小组在第二次世界大战期间进行的许多创新中的一个。第二次世界大战后，威尔克斯继续与军方合作，协助他们改善对装备进行的质量控制，利用统计方法发送对未来需求的计划工作，并把统计方法用于军事领域的所有方面。威尔克斯反对那些埋头于纯抽象理论的数学家，理由之一就是他们不爱国。他认为国家需要数学家的智慧，而这些人却精力用在没有什么价值的抽象世界里。国家需要这些人的智慧，以前是为了战时事务，后来则是为了冷战。

然而，没有记录显示有人曾对威尔克斯不满。他自由、亲切地面对每一位人，不管是刚踏出校园的毕业生，或是陆军的四星上将。他只是一个来自德克萨斯农场的老“小伙儿”，他会暗示对手，他知道自己还有很多东西需要学习，但他也想知道是否能……，接着是对所遇问题的详尽推理与分析。

抽象理论中的统计

威尔克斯尽力使数理统计不但成为数学里令人尊敬的一部分，还是一种实用的工具，他努力把同行的数学家们从冷酷的抽象世界中拉回来，不要为抽象而抽象。在抽象数学理论里，确实有一种基本的美感，这些形式上的美感如此吸引希腊哲学家柏拉图（Plato），以至于他声称，所有我们可以看到与接触到的东西，事实上只是真实世界的影子，而这个宇宙里真正能找到的真实事物，只能透过纯粹的理性来获得。柏拉图对数学的知识相当天真，其实希腊数学家所珍视的纯粹性，很多是有缺陷的。但是，透过纯粹的理性思考所发现到的美感，还是很诱人的。

自从威尔克斯成为《数理统计年报》的编辑之后，出现在该年报[29]和《生物统计》上的

[29] 20 世纪 80 年代早期，统计理论发展较快，因此，《数理统计年报》分成两种期刊：《统计年报》（Annals

文章越来越抽象。《美国统计学会期刊》（the Journal of the American Statistical Association）上的文章（这份期刊早期以政府统计项目为主）和《皇家统计学会期刊》上的文章也一样（早期刊登的文章不少是大英帝国的农业与经济统计方面的详细资料）。

曾经被数学家认为过度涉入实际问题泥沼的数理统计理论，此时已被重新澄清，恢复它的数学之美。通过高度抽象的理论归纳，亚伯拉罕·沃尔德（Abraham Wald）统一了已有的估计理论，被称为"决策理论"，在这种理论当中，不同的数理特性，会有不同的估计准则。费歇尔进行的实验设计研究，根据的斥是有限群论中的定理，用一些很巧妙的观点，比较不同的处理，由此推演出一个数学分支，称为"实验设计"（design of experiments）。但是，该领域的论文谈到的实验都较为复杂，因此，从未有实验科学家做过这种实验。

最后，当其他人继续研究安德烈·柯尔莫哥洛夫的早期著述时，概率空间与随机过程的概念变得越来越统一，但也越来越抽象。到了 20 世纪 60 年代，统计学期刊上的论文处理关于无穷集（infinite sets）的问题，通过对无穷集做并和交形成了西格互域（sigma fields）的集，即西格互域嵌套在西格互域中，使得无限序列在无穷远点收敛，而随机过程通过时间受限于一个小的有界状态集里，注定会永无止境地循环下去。数学统计的末世学，就和任何一种宗教的末世学一样复杂，甚至更复杂。数理统计的结论不但为真，更是可以证明其为真，这一点与宗教上的真理不所不同。

20 世纪 80 年代。数理统计学家认识到他们所从事的研究领域与现实脱离太远。为了满足应用的迫切需求，美国各大学纷纷成立应用性院系，如生物统计系、流行病学系、应用统计系等，设法调整这种分裂，它们原本属于同一学科。数理统计研究院（the Institute of Mathematical Statistics）的一些会议，冠上了"应用"的名义。《美国统计学会期刊》也另辟专栏，刊载相关的应用性问题，皇家统计学会的三份期刊当中，有份就命名为《应用统计》（Applied Statistics）[30]。但是，抽象理论的魅力仍在。成立于 20 世纪 50 年代的生物统计学，创办了《生物统计学》，打算刊登已经不受《生物统计》欢迎的应用性论文，但到了 80 年代，《生物统计学》的内容开始变的非常抽象，因此，又出现了其它期刊，如《医学统计》（Statistics in Medicine），以满足刊登应用性论文的需要。

当数理统计出现时，欧美各大学的数学系错失了发展良机。后来，在威尔克斯的带领下，很多大学成立了独立的统计学系。当数字计算机出现的时候，数学系很轻蔑地认为它只是一种从事工程运算的机器，又失去了机会。于是独立的计算机科学系成立了，有的从工程系分支出来，有的从统计学系分支出来。下一次重大革命是 80 年代分子生物学的发展，它牵涉到许多新的数学观点。正如第 28 章将会讲到的那样，数学系与统计学系都没搭上这班车。

威尔克斯逝世于 1964 年，享年 58 岁。在过去的 50 年间，他的很多学生都在统计学科的发展上发挥了重要作用。美国统计学会用他的名字成立了"S·S·威尔克斯将"（S. S. Wilks Medal），每年颁发一次，得奖人必须符合威尔克斯的数学创造力标准，以及对"现实世界"（real world）的热心投入。来自德克萨斯州的农家小伙，创造了自己的名声。

---

of Statistics）与《概率年报》（Annals of Probability）。

[30] 第二次世界大战结束后不久，《皇家统计学会期刊》一分为三，开始的时候名字分别为 JRSS-A、JRSS-B 及 JRSS-C，其中的 JRSS-C 最后改名为《应用统计》（Applied Statistics）。皇家统计学会的意图是，系列 A 的内容是与商业统计、政府统计有关的一般性问题。系列 B 的内容则全为数理统计，以及所有相关的抽象理论。但要维持《应用统计》的应用性是很困难的。每一斯当中都有几篇文章冠上"应用"二字十分牵强，说空了，只是在那里展示，又出现了一块美丽的抽象的数学宝石。

第 21 章 家庭中的天才

20 世界的前 25 年，数百万的移民从东欧、南欧迁往英国、美国、澳大利亚和南非。这些移民中的大多数来自他们本国的贫穷阶层，他们逃离压迫人的统计者和混乱的政府，寻求经济机会和政治自由。他们大都寄住在大城市的贫民窟，在那里，他们希望通过教育这个魔杖，使自己的孩子摆脱贫困。在这些孩子当中，有些人显示出非同寻常的潜力，有的甚至是天才。本章就介绍两个移民孩子的故事，其中一个拿到两个理学博士和一个哲学博士学位，而另一个，14 岁时就离开了就读的高中。

I·J·古德（I. J. Good）

古达克（Goodack）出生在波兰，但他不喜欢沙皇，也不喜欢沙皇对波兰的统治，特别不愿加入沙皇的军队。在他 17 岁的时候，就同与他有相同想法的朋友一起逃往了西方。他和他的朋友两人一共只有 35 卢布和一大块奶酪。一路上，他们没有车票，被发现时就用奶酪贿赂查票人员，晚上就睡在火车的座椅下面。古达克到达伦敦后，栖身在白教堂（Whitechapel）的犹太人贫民窟里，除了勇气和健康的身体外，当时他一无所有。后来他开了家修表店，而所有的修表技术都来自别的修表匠，他是靠在人家橱窗外偷看学会的（那里的光线倒很不错）。后来，他又对浮雕古董产生了兴趣，最后终于在大英博物馆附近开了一家古董珠宝店（从他未婚妻那里借的钱）。开业前，他雇了个画家，让他把自己的名字喷在新店铺的玻璃橱窗上，但那个家伙喝醉了酒，根本拼不出“Goodack”这几个字母，结果店名成了“古德浮雕定石之家”（Good’sCameo Corner），而这家人的姓氏也从此变成了“古德”。

古达克的儿子 I·J·古德 1916 年 12 月 9 日出生于伦敦。最初，古达克为儿子取名为伊西多尔（Isidore），但有一年，由于戏剧《善良的伊多西尔》（The Virtuous Isidore）到镇上演出，到处都张贴着宣传演出的大型海报，使年轻的古德非常尴尬。从那以后，他改名为杰克（Jack），并以 I·J·古德的名字发表论文和著作。

1993 年，在与大卫·班克斯（David Banks）的一次访谈中，杰克·古德回忆起他大约 9 岁的时候发现了数字的奥秘，并且心算能力变强。当时古德患白喉不得不卧床休息，他的一个姐姐来教他如何算平方根。在那里的正规学校课程安排中，学生学完长除法后，才开始学开平方，开平方的过程包含一连串的平分及平方运算，写在纸上有点像长除法的形式。

因为被迫在床上静养，古德开始用心算的方法开 2 的平方根。他发现计算好像可以一直延续下去，而且当他把已计算部分的结果再平方时，得数只比 2 小一点点。他继续心算下去，想看看能否找到某些模式或规律，但没有找到。他认识到整个过程可以看成一个数的平方与另一个数的平方的两倍之差，因此，只有当一定的模式存在时，这个数才可以用两数的比来表示。躺在床上，只靠心算，10 岁的古德就发现了 2 的平方根是无理数。与此同时，他也发现了“丢番图”（Diophantine）的问题的解，即“佩尔方程式”（Pell’s equation）。虽然早在古希腊时代，毕达哥拉斯学派（Pythagorean Brotherhood）就发现了 2 的平方根是无理数，佩尔议程式也在 16 世纪就解出来了，但这些都不影响一个 10 岁孩子在心算上的惊人成就。

在 1993 年的访谈中，古德沉思道：“那是一个不错的发现——曾被哈代（Hardy，活跃在 20 世纪 20－30 年代的英国数学家）称为古希腊数学家最伟大的成就之一。如果这一发现是当今的大人物所为，我会觉得很平常，但这在两千五百年前却是一个惊人之举。”

在 12 岁的时候，古德进入由缝纫用品商公司开办的艾斯克（Aske）男子中学[31]就读。这所学校位于哈姆斯代德（Hampstead），是专门为商贩的孩子们开办的学校，校规一向非常严格，它的校训就是要学会服务和服从（serve and obey）。在就读的所有学生中，大约只有十分之一能够升到最高年级；而这十分之一当中，又只有六分之一最后能进大学。在早年的求学生涯里，古德的老师是斯马特（Smart）先生。斯马特先生经常在黑板上抄一组练习题让学生去做，其中有些题是非常难的，他知道这要耗费学生很多时间，这样一来，他就可以利用这段时间在讲桌上做自己的事情。有一次，当他刚写完最后一题时，小古德就举手说："我做完了。"斯马特先生略带惊讶地问："你做完第一题了吗？""不！"古德回答："我全部都做完了。"

那时候，古德对数学难题的书异常地着迷。他喜欢先看答案，然后再在题目与答案之间找出一条捷径。在面对"一堆弹子"的问题时，他一看答案，就知道可以用比较繁琐的计算方法求出问题的解来。但对他来说，他感兴趣的是探索如果归纳解题的方法。在这个过程中，他发现了数学归纳法的原理，并完善了它。而这个原是仅仅是在 300 年前才被早期的数学家所发现。

19 岁的时候，古德进入剑桥大学。在此之前，有关他的数学天才的传闻，却比他的人更早传到那里。尽管如此，他还是发现，在剑桥有许多同学和他一样具有数学天分。那时候，剑桥耶稣学院（Jesus College）的数学导师似乎更喜欢规范的数学证明方式，以至于在整个数学证明过程中，任何直觉的思维成分，都要受到排斥。更糟糕的是，导师在黑板上写证明过程时的速度非常快，往往学生还来不及抄下来，就已经被擦掉，又写上了新了内容。古德在剑桥表现杰出，连一些资深的数学家都对他特别青睐。1941 年，他获得数学博士学位，论文阐述拓扑学的偏维（partial dimension）理论，是对亨利·勒贝格（就是前面曾提到过的那个成就令奈曼敬仰，但初次见面却对这个年轻人异常粗鲁的数学家）思想的扩展。

二战期间，古德成为一名密码破译员，他工作的地方就在伦敦附近的布莱奇利公园（Bletchley Park）里的一个实验室，其工作就是破译德国人的密码情报。一组密码往往由表述信息的字母转换成的一连串的符号或数字构成。在 1940 年，这些密码已变得非常复杂，转换的模式甚至可以随着每个字母的不同而改变。例如把"战争开始了"（war has begun）这段信息编成密码，一种方法是将这段话的每个字母配上一组数字，这样就构成了由 12 <u>06</u> 14 09 <u>06</u> 23 11 19 20 01 13 这样一行数字组成的密码。破译人员会注意到，其中 06 这组数是重复出现的，从而是可以判断它代表着同一个字母。如果这段信息足够长，且大约知道不同字母在语句中出现的统计频率，再加上一点幸运的猜测，密码破译员就有可能在几小时内把这段情报破解出来。

在第一次世界大战的最后几年，德国人研制出一种编码机器，可以为每个字母变换密码。譬如，第一个字母的编码也许是 12，而当这个字母第二次出现时，机器就会给它一个与上次完全不同的编码，这个字母的编码可能就变成了 14；等到第三次遇到同一字母时，也许编码又变了，如此这样编下去。依靠此种方法，密码专家就不会把上次已经使用过的数字，作为同一个字母的编码，再次使用。不过，作为密码的未来接收者，他们也必须了解这种新型密码的编制规律。因此，就机器编码来说，从一种编码转换为另一种编码，还是有一定的规律性的。密码破译专家可以依据一定的统计模式，估计出编码的规则性，从而找出破解密码的方法。然而，对于密码破译者来说，密码破译工作的难度还是越来越大：一旦最初的编码被一种固定程序所替换，那么整个程序就有可能被一种更高级的固定程序所替换，从而使衍生出来的新密码的破译难度更大。

[31] 缝纫用品商艾斯克学校是缝纫用品商公司创办的七所学校之一。缝纫用品商公司是一个古老的制服公司，成立于 1448 年，罗伯特·艾克斯，即公司原来的老板，死于 1689 年，死前留下了一个遗嘱：为可怜的缝纫商们的 20 个儿子创办一所学校。现在它是一所有 1300 名孩子的极为成功的学校，通过董事会，学校与缝纫用品商公司的联系仍然保持着，董事会的半数成员，包括主席，都是公司员工。

所有这些工作，都可以用一种数学模式来表示，它很像第 13 章里讲到的贝叶斯分层模型。编码的每一级的变换形式，都可以用一个参数来代表，因此，我们所面临的就是如何测量的问题：编码资料里的数字可当成观测的初始值，参数代表第一层编码，超参数描述参数的改变，超超参数代表超参数的变换，如此一层层下去。最后。由于密码总要被接收者破译，因此，到最后一层，此时的参数是固定不变的，所以理论上这种密码也是可以破解的。

古德的一项主要成就，就是他从做密码分析师的工作发展出来的经验贝叶斯法（empirical Bayes）与层次贝叶斯模型（hierarchal Bayes methods）。由于战争时的工作经验，使他对数理统计的基础理论产生了极大兴趣。后来他在曼彻斯特大学（University of Manchester）教了一段时间的书，但英国政府又诱劝他回到情报单位工作，在这里，他成为电脑处理分析密码的重要人物。电脑可以大量检验各种数字的可能组合，使他有机会研究分组理论（classification theory），在分组理论中，观察单位按“贴近度”（closeness）的不同定义来组织。

在英国情报单位工作的同时，古德又拿到两个更高的学位，即剑桥与牛津两所大学的理学博士。他 1967 年到美国，被维吉尼亚理工学院（Virginia Polytechnic Institute）聘为大学杰出教授，一直到 1994 年退休。

古德永远对偶然出现的数字巧合感兴趣。“我在本世纪第七个十年的第七年、第七个月的第七日的第七时，抵达（维吉尼亚州的）布莱克斯堡（Blacksturg），被安顿在第七街区的七号公寓）一切就是这么巧合。”接着，他又说：“我有个不太成熟的想法，上帝对那些愈不相信他存在的人，提供的巧合愈多。让这些人自己相信比强迫他们相信要好得多。”这双能发现数字巧合的眼睛，也瞄上了统计估计理论中的工作。由于人类的眼睛可以在纯随机的数字中，看出某些模式，因此他会问，这样一个明显的模式，它的真实意义是什么？古德用他的头脑，探索出了数理统计模型的根本意义，正因如此，他后来所写的论文和书籍，哲学的味道愈来愈浓。

迪亚科尼斯

佩尔西 • 迪亚科尼斯（Persi Diaconis）是希腊移民的后代，1945 年 1 月 31 日生于纽约。他的经历与 I •J •古德完全不同，但和古德一样，他从小就喜欢数学谜题。古德看的是 H •E •迪德内（H. E. Dudeney）写的书，书的内容在整个维多利亚时代的英格兰都很盛行；而佩尔西 •迪亚科尼斯读的是马丁 • 加德纳（Martin Gardner）为《科学美国人》（Scientific American）杂志撰写的“数学娱乐”（Mathematical Recreations）专栏。后来还是在高中的时候，迪亚科尼斯遇到了加德纳，加德纳的专栏经常介绍一些玩扑克牌的小把戏，和一些使事情看起来很不同的方法，这些都使佩尔西 • 迪亚科尼斯非常着迷，尤其是有关概率的复杂问题。

由于佩尔西 • 迪亚科尼斯太沉迷于扑克片游戏，因此 14 岁时就离家四处游荡。其实早在他 5 岁时，就表演一些魔术游戏。在纽约，他经常到一些魔术师聚焦的饭店或商店去。在一家餐饮他碰到了魔术师迪亚 • 弗农（Dia Vernon），弗农在全国各地旅行，表演魔术。弗农邀请他当助手，一起旅行表演。“机会来了。”佩尔西 • 迪亚科尼斯叙述到，“马上出发。我没有跟父母说一声，就跟弗农走了。”

当时弗农已经 60 多岁了，佩尔西 • 迪亚科尼斯跟了他两年，把他的道具与技术都学到手了。后来弗农在洛杉机安顿下来，开了一家魔术道具店，佩尔西 • 迪亚科尼斯继续一个人旅行表演魔术。别人觉得他的姓氏拼写比较麻烦，因此他给自己取了个艺名佩尔西 • 沃伦（Persi Warren）。就像他回忆的那样：

> 那并不是什么了不起的生活，但日子过得还不错。有一次，我在卡兹奇（Catskill）表演，有人看了我的表演之后觉得很喜欢，就过来说：“喂，老兄，想不想到波士顿表

演？……我可以付你 200 元美金。”……然后我就去了波士顿……安顿好表演场地，按确定的表演日期表演，……这时，或许就有经纪人来邀请你到别处去表演，日子就像这样。

24 岁的时候，迪亚科尼斯厌倦了旅行表演的生活，回到纽约。但他没有高中文凭。他原本在学校念书的时候还曾跳级，但 14 岁离家出走时，还差一年高中才毕业。由于没有高中文凭，他注册念纽约市立学院（City College of New York）的成人教育班。后来他发现在他离家的这些年里，许多军队和大学与理工学院都寄信给他，请他去读书，而且信的开头都称呼他为“亲爱的毕业生”。看来在他离家逃学之后，学校的老师决定无论如何还是让他毕业，因此把最后一年的分数也给了他，使他能顺利拿到毕业证书。迪亚科尼斯并不知道，其实他已经是纽约华盛顿高中（Washington High School）的正式毕业生了。

迪亚科尼斯上大学的理由很奇怪。他曾经买过一本研究生程度的概率论教科书《概率论导论及其应用Ⅰ》(Introduction to Probability Theory and Its Application, Vo1.Ⅰ)，作者是普林斯顿大学的威廉·费勒（William Feller）教授。他发现要看懂这本书很难（想看懂费勒这本书的大部分人都这样认为[32]）迪亚科尼斯进入纽约市立学院，想学到足够多的数学理论，以便把费勒搞懂。1971 年，他 26 岁时拿到了纽约市立学院的学士学位。

有好几个大学的数学研究生院都接受了他的就读申请，以前有人告诉他，哈佛大学数学系从没收过纽约市立学院的毕业生（其实是误传），因此他决定申请哈佛的统计系而不是数学系。他想去哈佛，他认为，进入哈佛后，如果自己不喜欢统计，“那我可以转念数学或其他学科。他们会知道我很棒……”因此会接受他转系。结果，他对统计很感兴趣，在 1974 年拿到数理统计博士学位，并接受斯坦福大学的一个职位，还一直升至教授。写本书时，他是哈佛大学的教授。

电脑完全改变了统计分析特性的结构。开始，它用来做费歇尔、耶茨及其他统计学家做过的同样类型的分析工作，只不过快得多，能量也大得多。还记得（在第 17 章）杰里·科恩菲尔德要算一个 24 阶矩阵的逆矩阵时遇到的困难吗？现在我桌子上的电脑可以算出 100 阶矩阵的逆矩阵（尽管总是碰到这种情形的人大概没有很好地定义问题），就连一些条件不够充分的矩阵，也能通过去处，求出广义的逆矩阵，这在 20 世纪 50 年代还只是纯理论的概念。对于实验设计产生的数据（涉及多重处理与交叉对照），大量复杂的变异分析都可以通过电脑来完全，这类工作涉及到的数学模型和统计观念，其实可以追溯到 1920 年到 1930 年。试问，电脑还有什么不能做吗？

在 20 世纪 70 年代，迪亚科尼斯和一些年轻的统计学家在斯坦福成立了一个研究小组，试图研究电脑和数理统计的结构，设法回答上述问题。他们最早提出的答案之一是“投影追踪”（projection pursuit）数据分析法。现代电脑带来的其中一项弊端，就是很可能组成一些难度庞大的数据组，假设我们正在跟踪一群经诊断为高危心脏病的病人，他们每半年到医院检查一次，检查时，每个病人抽取 10 毫升的血，分析血液中 100 种不同酶的尝试，其中有许多种被认为心脏病有关。此外我们为病人做心电图检查，测量六种不同的项目，并进行心电图监控（或者要求他们一整天都载着监控器，记录一天下来约 90 万次的心跳）。为了医疗诊断，该测的测了，该量的量了，该抽的也抽了，得到了 30-40 个测量结果。

怎么处理这些数据呢？

假设每位病人每次检查会产生 500 个测量值，而在研究期间必须跟踪 10 次，一个病人就有 5000 个测量值。如果总共研究 2 万个病人，可以描绘成一个 5000 维空间里的 2 万个点。通常在科幻小说里，仅有四维空间就可让人晕头转向，但在统计分析的真实世界里，处理数千维空间则是很平常的事。在 1950 年，理查德·贝尔曼（Richard Bellman）就提出了一组

[32] 数学教科书里，总有与书名相反的内容。最困难的书通常冠以“……简介”或“……原理”这样的名称，费勒的书困难程度应该加倍，因为它既是“简介”，并且又是第一卷。

定理，他把这组定理称为“维度的诅咒”（curses of dimensionality）。这组定理表示，当空间的维度增加时，得到确切参数估计的可能性就越来越小。一旦分析空间维度达到 10 至 20 个，观测值又少于 10 万，那么就分析不出任何结果。

贝尔曼的定理是基于标准的统计分析方法论。但斯坦福的研究小组发现，在这个 5000 维的空间里，这些真实的数据并非分散分布，实际上趋向较低的维度空间。假设这些分散在三维空间的点，全都落在同一个平面甚至同一条线上，这正是真实数据呈现的状态。每个临床研究病人的 5000 个观测值，不会毫无关联的呈分散状态，因为其中很多的测量值是彼此相关的。（普林斯顿大学和贝尔实验室的约翰·图基也曾提出过这种看法，他们认为至少在医学研究上，数据的真正“维度”通常不会超过 5。）根据这种思想，斯坦福研究小组发展了一种电脑应用技术，以找出实际存在的低维度空间。这些技术应用最广的就是“投影追踪”。

在此期间，由于大量的无序信息的增加，引起了其他科学家的注意，许多大学纷纷设立信息科学这门新科学。由于这些受过工程训练的信息科学家并不知道数理统计界的最新发展，因此会在计算机科学领域做平行发展，因而有时会重新发现一些统计学上已经知道的事，但有时也会打开一个全新的、费歇尔或他的追随者不曾预料过的领域。本书的最后一章，还会讨论这个问题。

第 22 章 统计学界的毕加索

我在 1966 年完成博士论文后，曾经拜访过一些大学，介绍我的研究成果，看看是否能找到一份工作。我的第一站是普林斯顿大学，当时约翰·图基亲自到火车站来接的我。

在我求学期间，就已听说过关于图基学术上的传说，图基单自由度交互效应（Tukey's one degree of freedom for interaction）、图基快速傅立叶变换（Tukey's fast Fourier transform）、图基快速检验（Tukey's quick test）以及图基引理（Tukey's lemma）。这些还不包括他在探索性数据分析（exploratory data analysis）研究中的成就和他在此后年代中的杰出贡献。图基是统计系主任（同时也供职于贝尔实验室），他亲自到火车站接我，使我受宠若惊。那天图基穿棉织长裤和休闲运动衫，脚上是一双运动鞋，而我却是西装革履。60 年代时尚风潮还没在大学教师中兴起，所以我的着装风格比他更正式。

图基带我穿过校园。路上我们谈论了在普林斯顿的生活条件，他还询问了我做论文时所用的电脑程序，他告诉我一些技巧，以避免程序中取整数上的差错。最后终于来到我要演讲的大厅。他把我介绍给大家后，就爬上了大厅的最后一排坐下。我开始演说，同时注意到，他正忙于修改学生的报告。

我讲完之后，有几个听众（都是研究生或教员）问了一些问题，并对一些细节提出建议。当确定没有人提问或评论时，图基就从后排走下来。他拿起粉笔，在黑板上把我的主要定理重写一遍，并且完全用我的符号[33]，然后用另一种方法，很快证明出这个花了我几个月才证明出的定理。“哇！”我对自己说，“真不愧为是大师！”

图基 1915 年生于马萨诸塞州的新贝德福德（New Bedford），他那特有的拖长声的波士顿近郊口音，使他的谈话更加风趣。他的父母在他很小的时候就发现了他的过人天赋，因此把他留在身边自己教他，直到图基进入布朗大学（Brown University）。在布朗大学他拿到了化学学士与硕士学位，但后来他被抽象数学所吸引，因此到普林斯顿大学继续研修数学，于 1939 年获得数学博士学位。他最初的研究领域是拓扑学（topology）。点集拓扑学是数学根本理论产生的基础，而在拓扑学的基础之下，是一个艰深而神秘的哲学支派，称为“哲理数学（或元数学）”（metamathematics）。元数学告诉我们数学问题的解意味着什么，在逻辑应用背后有哪些未明确的假设。图基深入研究这些混沌不清的领域之后，提出了图基引理，成为他在这个领域的主要贡献。

然而图基的学术归宿并不是抽象数学。普林斯顿大学的塞缪尔·S·威尔克斯教授，一直推动那些学生和年轻教员进入数理统计领域。拿到博士学位后图基留在数学系当讲师。1938 年，图基在准备论文时发表的第一篇文章就是有关数理统计方面的。后来到了 1944 年，他发表的所有论文几乎都是数理统计领域的。

二次大战期间，图基加入武器控制研究办公室（Fire Control Research Office）研究枪炮的瞄准、测距仪等与枪炮有关的问题。这种工作经历使他接触到许多统计问题的实例，

---

[33] 数学符号是由一串弯曲的希腊和罗马字组成，并且右上方和右下脚还加了一些字母或数字，常使外行人害怕（有时连数学家自己都头疼）。但数学符号又非常方便，它可以在有限的空间里，把许多复杂的概念连接在一起。阅读数学论文的“秘诀”，就是应该意识到每个符号都有特定的意义，在符号引入时就明确它的涵义，你还要坚信自己“已了解”它的意义，然后把注意力集中在符号的运用上。数学的精美之处就在于符号，它可以用简单的组合方式，使读者立刻明白不同概念之间的关系。我们在耶日·奈曼的论文中就能发现这种美，但我的博士论文就差的很远。我使用符号的目的，是要确保所有可能的数学模型牲都包括在内；我用的下标本身还有下标，上标也有自己的上标。有些地方我的下标有些含混，令我吃惊的是，那天下午第一次看到我的定理论证后，图基竟然就能全都记在脑海里，并把这一堆复杂的符号复述出来。（尽管我的数学符号有些混乱，图基还是决定给我一份工作，但我当时已有三个小孩，第四个也即将出生，因此，我最后接受了另一份薪水更高的工作。）

成为他后来研究的题材，也使他对实践问题的本质有了进一步的认识。他常用精辟的格言总结重要的经验，其中有一句来自他的实际工作，那就是："对正确问题的近似答案，胜过对错的问题的精确答案。"

多才多艺的图基

20 世纪初，出现了一位震惊世界的绘画大师 P·毕加索（Pablo Picasso），他的作品风格变化多端。有一段时间，他只用单色绘画，接着他又创造出立体主义，随后他又尝试古典主义形式，然后又去搞雕塑。毕加索每次的风格变化，都对艺术界造成革命性的影响，而其他人只能跟在他的后面，开发他留下恶报东西。图基也是如此。他从 50 年代开始研究安德烈·柯尔莫哥洛夫的随机过程概念，并发明了一种以电脑为基础的数据分析方法，可以分析一长串相互关联因素的影响结果，被称为"快速傅立叶变换"。就像毕加索的立体主义一样，图基在自然科学领域的影响是无人可比的。

在 1945 年，图基有关武器的研究把他带到了贝尔实验室设在新泽西州默里丘（Murray Hill）的研究中心，在那里他涉及到了各种不同的实际问题。在 1987 年的一次访谈中，他说："我们有位姓布登博姆（Budenbom）的工程师，他造出了一种新奇的雷达跟踪仪，可以用来锁定飞行目标。他希望能到加利福尼亚去发表一篇论文，为此他希望有一份能显示新仪食品跟踪误差的图表。"布登博姆以频率范围来表述他的问题，但不知道如何得到频率振幅的一致估计值。尽管图基作为数学家很熟悉傅立叶变换，但从未把这种技术运用于工程中。最后，图基提出了一个似乎能满足布登博姆需要的方法（还记得他的格言吗？正确问题的近似答案也是有用的）。但他自己对此方法并不满意，于是他继续思考这个问题）。

结果是快速傅立叶变换。他是一种修匀方法，用图基的话说，就是向邻近的频率"借力"，这样即使没有大量的数据，也可得到良好的估计值。此外，快速傅立叶变换也是一种经过慎重思考的理论解决方案，带有最适的特性。50－60 年代，在电脑的速度很慢、内存也很小的情况下，快速傅立叶变换还是一种非常有效的电脑演算方法。进入 21 世纪，这种演算方法依然有用，因为它比用更复杂的变换所得的估计值更精确。

电脑及其能力不断把统计研究的边界向前推进。我们在前面已提到电脑可计算大型逆矩阵的能力（这些如果让约翰·科恩菲尔德（John Cornfield）用手摇计算机做，可能需要数百年时间），此外，电脑在统计理论上还有另一压倒性优势，就是电脑的储存与分析大量数据的能力。

在 60 年代与 70 年代早期，贝尔实验室的工程师和统计学家是分析大量数据的先驱。监视电话线路的随机误差和问题，导致成千上万的数据项都存在一个电脑文件中，而用太空探测器传回的火星、木星及其他行星的数据资料，项目也都是数百万笔。你要如何看待如此大量的数据？又要如何整理它，才能加以检验？

按照 K·皮尔逊开创的方法，我们总能估计出概率分布的参数，这就需要我们对这些分布做些假设，比方说假设这些分布属于皮尔逊系统。但如果我们不对分布做特别的假设，能不能有方法检验大量的调查数据，得到我们所需的信息呢？从某种意义上说，优秀的科学家一直是这么做的。格雷戈尔·门德尔（Gregor Mendel，奥地利遗传学家）做了一系列植物杂交实验，检验得出的实验结果，逐渐发展出他的显性和隐性基因理论。虽然大量的科学研究涉及到收集数据，并把收集到的数据和预先存在的某种分布模型对比，但有时仅收集数据，仔细地加以检验以发现意外结果也是非常重要和有意义的。

正如美国数学家埃里克·坦普尔·贝尔（Eric Temple Bell）曾经说过的："数字不会说谎，但它有个偏好，就是在存心说谎的时候讲出真相。[34]"人类倾向于寻求模式，并往往

[34] 贝尔在 1940－1950 年写了几本数学普及性读物，他的《数学人》（Men of Mathematics）至今仍是描写

在只有一些随机的、模糊的信息时，就认为已经找到了模式。[35]

这种现象在流行病学中比较明显，我们在调查数据时，常常发现在某些地方或某些时段有些疾病容易“群发”。假设我们发现马萨诸塞州的某个小镇，儿童患白血病的人数异常偏高，是否表示该镇上存在某种致癌因素？或者这只是碰巧发生的随机群体，在其他任何地方也有可能发生？假设当地居民发现有化工厂往镇的湖里排放化学废弃物，假设他们同样发现在儿童患白血病例较多的地区，饮水中芳香族胺（aromatic amines）的尝试较高，我们是否可以断定这就是导致儿童患白血病的原因呢？从更广义上说，在多大程度上，我们可以用倾向于模式的目光去检验数据，并且可以期望找到比这些随机的、模糊的讯号更多的信息？

在 60 年代，图基开始认真地考虑这些问题。他从这些问题中发现一种数据处理方法，可以说是 K·皮尔逊方法的精炼版本。他认识到，即使没有武断的概率模型设定，还是可以把观测数据的分布当作一个分布来检验。结果，他发现了一系列论文，参加了很多场演讲，最后写成了几本书，被称之为“探索性数据分析”（exploratory data analysis）。在处理这些问题的过程中，图基采用了一种十分原始的方式来阐述他的观点。为了引起他的听众和读者的注意，使他们重新检验相关的假设，他对以前使用过的数据分布特征重新命名。同样，他脱离以往用标准概率分布的这个分析起点，转向检验数据本身的模式或形态，他还审视极值能改变我们观察模式的方式。为了调整错误的印象，他发展出一套图形工具来显示数据。

例如，他指出我们常用来表示数据分布的直方图（histograms），容易给人造成误导，会引导观测者去注意那些频繁出现的观测值。因此，他建议以观测值次数的“平方根”（square root）来观测值出现的次数，并以此数据画出的图形来取代直方图。他称这种图为“根图”（rootgram）。图基还建议将数据分布的中央区域画成一个小盒子开关，而把极值画成由盒子延伸出去的线段（他称这些线段为“胡须”（whiskers））。他提议的统计工具，有许多都被纳入标准的统计软件包。现在的分析师称它们为“箱形图”（box plots）和“茎叶图”（stem and leaf plots）。图基丰富的想象力扫遍整个数据分析领域，他的许多建议至今还在电脑软件中应用。我们至今用的两个英文单词，bit（位或二进位）和 software（电脑程序，相对于电脑硬件）就是图基创造的。

对图基来说，世上没有什么事情会因为平凡而不值得去发挥原创力，也没有什么事情神圣到不容质疑。就拿最简单的记数过程来说：许多读者在计数某种东西时，或许已使用过一种记数符号。一代代的老师教我们的常用的符号就是先画 4 条垂直竖短线，第五条线穿过这 4 条线，表示 5 个数。不知读者看到过多少这样的场景：衣衫褴褛的犯人在监狱的墙上画下了一串串这样的计数符号。

图基说，这其实是一个愚蠢的记数方法。想想看，它多么容易出错。你可能画了三条竖线就画一个横线，也可能画了五条竖线后才画横线，这种记数法即使错了也很难发现，除非你仔细检查所画垂直线的数量。用一种容易找到误差的记数符号似乎更有意义。图基提出了十笔记数法：首先画四个点作为方型的四个角，然后再把四个点连成四条线，形成一个方型，最后在方型内画两条对角线。画完之后是十笔。

上述这此例子，快速傅立叶变换、探索性数据分析，都只是图基巨大成就的一部分。就像毕加索从立体主义到古典主义，从雕塑再到建筑，图基在 20 世纪下半叶，畅游于统计学的各领域，从时间序列（time series）、线性模型（linear models），到费歇尔的一些被人

18、19 世纪伟大数学字的经典传记资料。他的《命理学》（Numerology），也就是这句话的出处，内容与命理有关，据他的作品记载，他是通过他的女佣的介绍才进入这个领域的。

[35] 最著名的例子是“博德定律”（Bode's law）。这是个观测得到的经验定律，是指在太阳系中，行星与太阳之间距离的对数值，与该行星在太阳系中的顺序成某种线性关系。实际上，海王星的发现就是依据博德定律，天文学家预测的轨道上寻找，结果真的找到了海王星。直到木星与土星太空探测器发现这两颗行星有许多更小的卫星之前，惟一被观察到的木星仍然符合博德定律。博德定律是偶然巧合？还是能真的告诉我们，太阳与行星之间有更深的、尚未被人们了解的关系？

遗忘的研究工作的推广，再进一步到稳健估计（robust estimation）及探索性数据分析。他从研究深奥的数学理论起家，又因思考和解决实际问题脱颖而出，最后落脚在研究无结构的数据估计上。在他研究的所到之处，统计变得与以往大不相同。就在 2000 年夏天，也就是在他去世的当天，他还和朋友、同事们在一起，讨论问题，提出自己的新观点，并对以往的旧观点展开质疑。

第 23 章 处理有瑕疵的数据

证明统计方法用途的数学定理通常都假设:在科学实验或观察中的测量值都是同样有效的。如果分析者在进行分析时，只选择数据中他认为看起来是正确的数据来分析，那么统计分析结果可能就会产生非常严重的错误。当然，这正是以前科学家们通常的做法。早在 20 世纪 80 年代初期，S •施蒂格勒阅读了 18 世纪和 19 世纪许多伟大科学家们的笔记本，比如，因为确定了光速而获得 1907 年诺贝尔奖的艾伯特 • 迈克逊（Albert Michelson）。施蒂格勒发现，所有这些科学家在开始他们的计算前已经剔除了一些数据，17 世纪初就发现行星绕太阳以椭圆轨道运行的科学家约翰尼斯 • 开普勒（Johannes Kepler），他在研究古希腊天文学家的记录时，发现有一些观测位置记录不符合他正在计算的椭圆轨道，于是他就忽略了这些缺损数据（faulty value）

但是现在，值得尊敬的科学家们不再抛弃那些看起来是错误的数据，统计革命在科学界的广泛影响，教会了现在的实验科学家们不要剔除任何数据。统计学的数学定理要求同等对待所有的数据。但如果有些数据的确错了，我们该怎么办？1972 年的一天，一位药理学家带着这样一个问题来到了我的办公室。他在小白鼠身上研究溃疡的预防，正在比较两种不同的处理方法，他确信这会产生截然不同的结果，而且他的数据看起来也显示同样的结论，但是当他依据奈曼－皮尔逊的理论进行正式的假设检验时，比较结果并不显著。他确信问题出在两只小白鼠的观测数据上，这两只小白鼠使用了不足量药剂，尔后都没有发生溃疡，使得它们的结果看起来要远远好于另外一种处理方法的实验结果——而那本应该是最好的。我们在第 16 章已经看到了非参数方法是如何发展起来去解决这一类问题的。这两个离散数据刚好处于错误的一边，而且数量上还是两项，所以即使用非参数检验结果也不显著。

如果这种事情发生在一百年前，这个药理学家就可以剔除这两个错误的数据，继续进行他的计算，不会有人提出异议。但是，他已经学习了现代统计方法，他知道他不能够这样做。很幸运，当时我手头正好有一本刚读过的新书，书名是《位置的稳健估计：调查与与改进》（Robust Estimates of Location: Survey and Advances），它记述了一项重大的主要应用计算机进行的研究成果，即约翰 •图基进行的我们称之为“普林斯顿稳健性研究”（Princeton Robustness Study），在这本书中我们可以找到这位药理学家问题的答案。

“稳健（robust）一词对很多美国人来说，听起来很奇怪。许多统计学术语都来自于英国的统计学家，并且都反映了他们的语言习惯。例如，在英国，把数字微小的随机波动称为“误差”（error）是很普遍的[36]，有时候，数据不仅是明显错误的，而且由这引动错误造成的结果的原因也是可能看出来的，例如一块田里的农作物绝产。这样的数据被费歇尔称为“谬误”（blunders）。

是乔治 • 博克斯（George Box）——费歇尔的女婿，在他的英国语言应用习惯的基础上发明了“稳健”（robust）这个词。博克斯有很得的口音，这主要是因为他最初成长在泰晤

[36] 当一份统计分析报告中出现同样的词汇时，我们经常搞不清楚他代表的是词汇的一般意义还是特殊的统计意义。当我最开始在医药行业工作时，我的一份分析报告中包括表明结果的传统表式，其中一行是非常小的随机影响而带来的不确定性，按传统的用法我称之为“误差”（error）。一位高级执行官拒绝将这份报告上报美国食品和药物管理局（FDA），“我们怎么能够允许在我们的数据中出现‘错误’呢？”他问，并要求进一步核实数据。其实我们已经做了相关的工作。我告诉他这只是传统的说法而已，但是他坚持要我找到其它方式来描述它，他不会将这份有“错误”的报告上报美国食品和药物局（FDA）。我联络了康涅狄格州大学的 H • F • 史密斯先生，把问题告诉了他，他建议我称之为“残差”（residual），他告诉我，在许多论文中它都被称为“残差”。后来，我对其他从事药业工作的统计学家提起这事，他们也开始使用这种称呼了。最后，它成了医学文献中标准的术语。看来，没有人，至少是在美国，是允许存在“错误”的。

士河附近。他的祖父当时是一个五金器具批发商，生意很不错，供博克斯的伯父们读完了大学，其中有一位还成了神学教授。当博克斯的父亲成年时，祖父的生意已经失败，他父亲没有受过高等教育，只好去作一个商店主的助理，靠薪水维持全家人的生活。博克斯上了中学，知道他没有钱上大学，所以他开始在一个技校里学习化学。这时，第二次世界大战爆发，博克斯应征入伍。

因为有学习化学的背景，他被分配去化学防御实验部门工作。在那里，许多顶尖的英国药理学家和生物学家正致力于不同毒气解毒方法的研究。约翰·加德姆爵士（Sir John Gaddum）也在这些科学家中，他在 20 世纪 20 年代末将统计革命引入药理学，并且为药理学的基本概念赋予了一个牢固的数学基础。

博克斯成为一个统计学家

博克斯的上司是一个陆军上校，他对收集来的大量数据感到束手无策，这些数据记录的是不同剂量的不同毒气在老鼠和小白鼠身上的不同反应。他搞不清楚这些数据说明了什么，就像博克斯在 1986 年叙述的那样：

> 有一天，我对长官说：“你知道，我们真的需要有个统计学家来帮我们看看这些数据，因为它们变化太多了。”他说：“是呀，我知道。但是我们找不到一个统计学家，因为它们都很忙。你对统计知道些什么？”我说：“噢，我对此一无所知，但是我曾经读过一本书叫《研究工作者的统计方法》，是一个叫费歇尔的人写的，我没看懂，但是我想我明白了他正在做什么。”于是长官说：“那好，如果你读了这本书，最好由你来做这件事吧。”

于是，博克斯与军队的教育机构联络，要求去进修统计方法的课程。但是当时没有这样的课程，统计分析方法还同有成为大学的正规课程，但是他们送给博克斯一份阅读书目，书目无外乎最新的图书出版信息，其中列有费歇尔写的两本书，一本关于教育研究的统计方法，另外一本关于医学统计学，此外，还有一本书是谈林业和牧场管理的。

博克斯对费歇尔的实验设计非常感兴趣。他在那本关于林业管理的书中发现了几个特别的设计，并将这些设计改造，使之适合于进行动物实验（当时科克伦和考克斯合著的《实验设计》一书尚未出版，书中有许多细心描述的实验设计）。通常由于书中所列的实验设计不是很适用，所以博克斯就参照费歇尔的一般性的描述，结合他的发现，考虑了自己的实验设计。其中有一个最让人感到奇怪的实验是：让志愿者两臂各露一小块皮肤，暴露在不同的毒气下，然后采用不同的治疗方法。每个人的两臂是相关的，因此在分析时必须考虑这个因素，必须做一些处理，但是在这本关于林业的书中没有这方面的论述，在费歇尔的书中也没有类似的论述。所以，博克斯这个只在技校里不完整地进修过一些化学课程的，只好从基本的数学原理开始，创造出适用的实验设计。

博克斯实验设计的实力在一个否定结论的实验中表现出来。一个美国眼科专家带着他认为对刘易士毒气（lewisite）治疗效果极好的解毒剂来到了博克斯的实验室。刘易士毒气毒性极强，一小滴就可导致失明。他在美国已经在兔子身上做了很多次试验，他的厚厚的论文也证明了他的药剂效果极好。当然，他根本不知道费歇尔的实验设计，事实上，在他的实验中漏洞百出，实验设计中有许多与结果无关的因素没有分离出来，这样的设计是不可能得到真实的结构的。兔子有两只眼睛，于是博克斯利用他的新设计针对这个事实提出了一个非常简单的实验，这个实验很快显示这种解毒剂根本是无效的。

他们准备写一份描述这些结论的报告，作者是一个英国军官，博克斯负责写统计附录，即解释这个结论是怎样得出的。一个负责审核报告的军官坚持删除博克斯写的那部分，他认为这部分太复杂了，没有人能看懂（事实上是这位负责审查的人看不懂）。但是约翰·加德

姆爵士已经阅读了初稿，他跑去恭贺博克斯在附录部分所做的工作，得知这部分将在最终报告中删除，于是他拉着博克斯怒气冲冲地闯进了组合行军棚屋，当时审查报告委员们正在开会，用博克斯的话说："我感到很尴尬，这个非常有名的大人物为在场的所有国家公职人员读了一段我写的附录，然后说：'把这些东西给我放回去'。"他们很快就照办了。

战争结束后，博克斯认为去学习统计学是非常有价值的，他已经读了费歇尔的书，知道费歇尔在伦敦大学的大学学院任教，于是他来到了这所大学，但是他不知道费歇尔已经在1943 年离开了伦敦大学到剑桥大学任遗传系主任了。会见博克斯的是 E·皮尔逊，费歇尔曾对他跟奈曼合作进行的假设检验进行过刻薄的批评。会谈时，博克斯热情洋溢地描述他对费歇尔理论的认识，介绍他在实验设计中的心得，皮尔逊静静地听着，最后说："好吧，总之你可以来我校就读，但是我想你将来会知道，在统计界里除了费歇尔外，还有其他一个或两个人的存在。"

博克斯留在大学学院里学习，取得了学士学位，接着又继续攻读硕士学位。他发表了许多关于实验设计的文章，被认为可以当作博士论文，于是，他直接得到了博士学位。当时，帝国化学工业公司（Imperial Chemicals Industry（ICI））是英国最主要的发明新化学药品的公司，博克斯应邀参加了该公司的数学服务小组，他从 1948 年至 1956 年一直在 ICI 公司工作，其间他写了一系列的论文（通常是合著），这些论文扩展了实验设计方法，检验了一些在生产过程中为提高效益进一步调整产出的方法，同时，也是他后来对柯尔莫哥洛夫随机理论进行应用研究的起点。

博克斯在美国

博克斯到了普林斯顿大学任统计方法研究小组的负责人，接着到威斯康星大学开设了统计学系。他已经是所有重要统计组织的成员，因为他卓越的成就得到了好几项声望很高的奖励。即使在退休后，他仍然致力于学术研究和学术组织的管理工作。他的研究成果覆盖了很多统计研究领域，不但有理论研究还有应用研究。

博克斯在帝国化学工业公司工作时认识了费歇尔，但是私交并不深。当他在普林斯顿大学负责统计方法研究小组的工作时，费歇尔的一个女儿琼（Joan）得到了一个去美国的机会，她的朋友为她在普林斯顿大学找到了一个秘书的工作，博克斯与她相遇，后来两人结了婚。琼在 1978 年时出版了一本权威性传记，记录了她父亲和她丈夫的工作。

博克斯还有一个对统计的贡献就是"稳健"（robust）一词。他考虑到很多统计方法都是依赖于数学定理的，而这些数学定理对数据分布特性的假设可能不正确，如果数学定理的条件不成立，能找到可用的统计方法吗？博克斯提议称这些方法为"稳健方法"。他做了一些初步的数学研究，发现"稳健性"（robustness）的含义太不明确，但他反对对此概念赋予更加明确的含义，因为他认为一个概括性的模糊思想会对方法的选择更加有利。然而，这种思想本身还是得到了发展，用一个术语定义假设检验的稳健性就是：误差概率（the probability of error）。斯坦福大学的统计学教授布拉德利·埃弗龙（Bradley Efron）把费歇尔的一个几何学概念作了延伸，他在 1968 年证明了"学生"t－检验具有稳健性，他还用 E·J·G·皮特曼（E. J. G. Pitman）的方法证明了大多数的非参数检验也是同样稳健的。

20 世纪 60 年代末，普林斯顿大学的图基和他的研究小组成员以及他的学生们，研究如何处理那些显而易见是错误的测量值。他们的成果就是 1972 年发表的"普林斯顿稳健性研究"。这项研究的基本观点是有瑕疵的分布（contaminated distribution）（*有的辞典上将之翻译为污染分布——译者注*）。通常情况下，我们假设取得的测量值绝大部分是来自于一个概率分布，而且这个概率分布的参数是我们要估计的，但是，测量值当中总会有极少的一

些测量从上到下为自于另外一个分布，所以我们说这些测量值是有瑕疵的。

在第二次世界大战期间，有一个典型的关于瑕疵分布的例子。美国海军改进了一种新型的光学测距仪，要求使用者用一个三维立体镜去看目标的影像，用一个大三角“罩”在目标上，为了确定这个仪器的统计误差，让几百名水手来试用，测量一个已知距离的目标。在试用前，根据随机数表重新确定了目标的位置，这样后来的水手就不会受先前已知位置的影响。

设计这个研究的工程师不知道，有 20%的人看东西不是立体的。因为他们是我们所说的弱视（lazy eye），这样有五分之一的数据是完全错误的。单从手头研究得到的数据看，不可能知道哪些数据是来自于弱视者的，因此分不出哪些数据来自于有瑕疵的分布。

普林斯顿的研究是在计算机上实施蒙特卡罗法（MonteCarlo）[37]模拟计算大量来自有瑕疵分布的数据，寻找估计这个分布的中心趋势的方法。当数据有瑕疵时，一般人通常喜欢用的平均数是不可靠的，关于这一点也有一个经典的例子，讲的是 20 世纪 50 年代耶鲁大学所做的一次试验，估计该校的毕业生 10 年后的收入情况。如果他们用平均值，那么收入是非常高的，因为有几个当时是千万富翁，但是，事实上，80%以上的毕业生平均收入均低于这个平均数。

“普林斯顿稳健性研究”发现，平均数在一个有瑕疵的分布中受个别值的影响往往很大，这正是那位药理学家告诉我的小白鼠溃疡研究实验中出现的数据问题，而这位药理学家所学的统计方法都是用平均值来做分析。读者可能会问：如果这些极端的、而且看起来是测量值实际上是对的，假设他们是属于我们正在检验的面盆，并不是来自另外的分布，会怎么样？如果将这些数据剔除，结论就会产生偏差。

普林斯顿的稳健性研究找到了一个解决方案，有以下两种方法：

1. 如果测量值有瑕疵，就降低瑕疵测量值的影响力；
2. 如果测量值没有瑕疵，就找出正确的答案。

我建议这个药理学家使用其中的一种方法，这样他就可以根据数据得出正确的结论。后来他的下一步实验得到了一致的结果，说明稳健分析是对的。

博克斯与考克斯

博克斯还在帝国化学工业公司工作的时候，他经常去拜访大学学院里的统计小组，在那里他遇到了大卫•考克斯。考克斯已经成为统计的主要创新者，是《生物统计》（K•皮尔逊的期刊）的主编。这两个人都觉得他们的姓氏相像，很有意思，而且博克斯和考克斯连起来刚好是英国戏剧里的一个术语，意思是一个赏扮演两个小角色，还是一个英国经典音乐讽刺喜剧中的两个人物的名称，剧中，博克斯和考克斯租住一间房里的同一张床，一个白天睡，一个晚上睡。

博克斯与考克斯决定共同写一篇论文。但是，他们在统计领域中的兴趣不同，随着时间的推移，他们一再地努力，但是他们的兴趣实在是太不相同了，这样，如果要共同写这篇论文，他们就不得不对各自持有的关于统计分析性质的不同角度进行调和。1964 年，他们的论文终于在《皇家统计学会期刊》上发表，就如这篇论文广为人知一样，“博克斯•考克斯”成为统计方法中的一个重要部分。在这篇论文中，他们阐述了如何用一种方法转换测量值，使得大部分的统计程序更具有稳健性。用他们的名字命名的“博克斯－考克斯变换”（Box-Cox transformations）方法用于研究化学物质使活细胞突变的效应，也用于经济计量分析，甚至用于农业研究——费歇尔方法最初产生的领域。

[37] 蒙特卡罗法中，每个测量值是用随机数模拟可能发生的真实事件产生的，这个模拟过程要做数万次，利用产生的值进行统计分析，确定特定的统计方法在模拟情况下的结果，这个名字源自于摩纳哥的一个著名的赌城。

第 24 章 重塑产业的人

1980 年，美国国家广播公司（NBC）播出了一部电视记录片，片名为《日本人能，我们为什么不能？》。美国汽车公司被来自日本的挑战震惊了：从 70 年代起，日本生产的汽车在品质上已远远超过了美国生产的汽车，但价格却比美国低得多。不仅是汽车，其它工业品，从钢铁到电子产品，日本和美国相比，在质量和价格上都占优势。NBC 的记录片就是要探讨这是怎么发生的。这部纪录片实际上推出一个人——时年 80 岁的美国统计学家 W·爱德华兹·戴明（W. Edwards Deming），是他影响了整个日本的产业界。

一时间，戴明成为美国产业界的热门人物。其实，戴明自奉人 1939 年离开美国农业部以来，一直在产业界从事咨询顾问的工作。在从事这一职业的岁月里，他曾多次受美国的一些汽车公司的邀请，协助他们进行质量管理工作。正是在这一长期过程中，戴明对如何改进产业形成一套有效的方法。但是，美国这些公司的高层管理者却普遍地认为，质量管理不过是些“技术性”的细节，对此他们没有兴趣。他们认为，进行质量管理，只需雇请一些专门人员就足矣。到了 1947 年，G·麦克阿瑟（G. Mac Arthur）将军被任命为日本占领区的联军最高怀念，他强迫日本政府采纳西方国家的民主宪政制度，并且召集了一批一流的专家来日本，以“美国方式”（American way）来教育这个国家。于是，他的手下将戴明以统计抽样专家的名义邀请到日本，教授日本人“美国人是怎么做的”。

戴明的课程深深打动了一个叫石川一郎（Ichiro Ishikawa）的日本人，所以，后来他作为日本科技与工程联合会（JUSE）的主席，再次邀请戴明来日本，在产业界的一系列研讨会上讲授统计方法。石川一郎在日本产业界很有感召力，在他的邀请下，许多高级管理人员也经常来听取戴明的讲课。在那个年代，“日本制造”这几个字，就是“廉价、粗制滥造的仿制品”的意思。在戴明的研讨会上，他大胆地告诉他的听众们，不出 5 年，这种善就可以改变。只要适当运用统计方法的质量控制，他们就能够生产出物美价廉的产品来，从而，他们将迅速占领世界各地的市场。戴明后来承认他所说的 5 年是低估了，日本人差不多只用了两年的时间就改变了他们的状况。

戴明的作为在日本的产业界产生了极其深远影响，为此，日本科技与工程联合会（JUSE）专门设立了一项以戴明的名字命名的年度奖，用以奖励产业界那些在质量管理方面做出杰出贡献的人。日本政府也看到了运用统计方法改进各项活动的前景。日本教育部还专门选择一天作为“统计日”（Statistics Day），在这天，学生们要开展统计知识创新展示的竞赛活动。总之，统计方法风行于全日本，这几乎全都来源于戴明的讲座。

戴明带给高级管理层的信息

1980 年 NBC 的电视记录片播出后，戴明的名字开始在美国产业界受到欢迎。他开办了一系列的讨论，传授自己的美国管理理念。不幸的是，大多数美国公司的高级管理者并不明白戴明所做的事。他们只是派出一些已经知道质量管理的技术专家来听戴明讲课，很少有来自公司高级管理层的主管人员出席。而戴明的讲座内容主要是针对企业的高级管理层的，其中充满批判精神，听起来让人感到有些刺耳、不愉快。管理层，尤其是高级管理层，没有做好自己的工作。为了能以实例阐述自己的观点，戴明特意邀请了一批学员参与他在制造业的一项实验活动。

参与实验的学员被分成工人、巡视员和管理者三组。工人们将被训练从事一种简单的生产程序。先发给他们每人一个大圆桶，桶里装满珠子，珠子以白色的为主，其中搀有少量红色的。首先，工人们要竭尽全力摇晃这个圆桶，以使里面的珠子分布均匀，他们被告知此举

是至关重要的一个环节。然后，发给他们每人一个木铲，木铲上面排列着 50 个小坑，每个坑的大小正好能放一颗珠子。要求工人们利用这个木铲从桶中取珠子，每次正好 50 颗。训练者告诉工人们，50 颗珠子中红色的至多不能超过 3 颗，否则在市场上顾客不会买账，他们必须想方设法达到这个目标。在整个实验过程中，每当一个工人取出珠子，巡视员就记录下其中红色珠子的数量，管理人员会检查记录，表扬那些做得好的——每次红色珠子少于 3 颗和正好等于 3 颗的工作：批评那些做得差的——每次红色珠子多于 3 颗 的工人。实验上，那些做得差的工人很可怜，时常被管理者要求停下手中的工作，去看那些做得好的工人是怎么做的，以学习他们正确的操作方法。

在这个实验中，所给的每个桶里的珠子中红色的数量约占 1/5。而在这样的条件下，要使每 50 颗珠子中红色珠子等于或少于 3 颗的机会还不到 1%；而获得 6 颗或 6 颗以下了机会是 10%。所以，工人们为了这个难以达到的目标——每 50 颗珠子中红色珠子正好等于或少于 3 颗而拼命努力。但实际上，平均来看，每次工人所取出的珠子中，约 10 颗是红色的，这是管理者所不能接受的；而按概率来看，有的工人甚至会取出 13 至 15 颗红珠子。显然是工作极差的结果。

戴明的观点是，通常情况下，管理者往往设立一些不可能实现的标准，他们不在意标准是否可以达到，也不尝试着如何通过发送设备等必要手段，来使这些标准得以实现。相反，美国企业中的高层管理者们，往往只是领先质量管理专家来制定的标准要求工人，而根本不管工人们会遭受的挫折。这种现象在当时成了美国产业界的一种通病。戴明对此提出尖锐批评。在 70 年代，风行于美国产业界的所谓“零缺陷”（zero defect）理论，其核心就是要求企业生产的产品没有任何缺陷，戴明认为这是根本不可能做到的。到了 80 年代，产业界又兴起所谓的“全面质量管理”（TQM）之风（此时正是戴明刚刚在美国产业界出名的时候），戴明指出，这全是些没用的空话。他劝告企业管理者们还是做点实事。

戴明在他所著《走出危机》（Out of the Crisis）一书中，引用了他写给某公司管理者的一份报告。报告指出：

> 本报告是应贵公司之邀在对贵公司目前的问题：产量下降，成本上升，产品质量不稳定……研究之后写成的。我们的看法是，除非公司高层负起责任，否则无法在改进质量上建立起永久的机制。在我看来，你们的麻烦的主要原因是，你们的管理层没有对质量负起应有的责任……，你们公司所具有的不是质量控制而是打阻击战的游击队，没有组织好的系统，没有预防措施，也没有把质量控制看作一个系统，你们经营的是一个消防队，只指望出事时及时到达以附上火热蔓延。
>
> 在你们公司里，到处都能看到一个鼓动性口号，号召每个人都要把工作完成的尽善尽美。但我想知道，你们究竟怎样做才能使每人实现这一目标。如果工人对自己的工作并不了解，不知道如何才能把工作做好，又怎么能做好呢？如果原材料的质量不合格，或供应不及时；机器设备出现了故障，又怎么能把工作做的完美？除此之外，另一个管理层认识的误区是：只要生产线上的工人按规定的要求去做，生产中就不会出现问题。所以一旦出现问题，就是生产工人的责任，而与管理者无关。
>
> 就我本人的经验来看，生产中出现的绝大多数问题都有其共同的原因，而只有管理层可以减少其影响，或将之根除。

戴明关于产品质量管理的主要观点是：产品的生产过程是可变的，之所以这么说，原因在于那是所有人类活动的特性。什么是消费者最希望的产品？对此问题，戴明强调：消费者最希望的产品并不是完美无缺的，而是质量稳定可靠的（reliable）。他们（她们）希望所购买的商品质量稳定，这样就可以从中得到消费预期。依据费歇尔的变异分析理论，生产过程中的变异有两方面的来源：一个原因戴明称之为特殊原因（special causes），另一个他称之为一般原因（common），也可称为环境原因（environmental）。戴明主张，美国产业界

应该制订相应产品生产的标准程序，允许产品生产过程在一定范围内变化。一旦生产过程中出现的问题超出这一界限，即停下来寻找问题出在何处。戴明指出，由于特殊原因导致的问题不多且很容易被发现。而环境原因的问题总是存在，这是管理不善的后果，它们通常以机器设备缺乏维修保养、原材料供应质量没有保障，工作条件失控等形式表现出来。

戴明指出：生产线就如同一条活动的河流，从原材料供应开始，到产成品产出，每一个环节都可以被测量。由于环境原因，每个环节都有其自身的变化。管理者不能坐等最终产品超出前定的变化范围，而应该密切注视每个环节的变化，变化最大的环节要作为控制的重点，一个环节的变化被减弱后，另一环节就会突显而成为新的重点。因此，质量管理是一个连续性的过程，生产线上最突出的问题始终要加以解决。

日本人在采用了戴明的方法后，其生产的汽车可以行驶 100 000 公里以上无大修；船只只需极少的维修；所生产的钢铁质量稳定，几乎每批都一样；其它工业产品的质量也都得到有效的控制。

质量管理的特性

从 1920-1930 年，贝尔实验室的沃尔特·休哈特（Walter Shewhart）和国家标准局的弗兰克·尤金（Frank Youden），组织了第一个统计质量管理计划，将统计革命引入美国产业界。戴明也积极鼓动将这场统计革命引入上层的管理部门。在其专门为管理者所著的《走出危机》一书中，戴明力图以最有限的数学知识，讲解有关的管理理念。他指出了制造业中普遍存在着的糊涂观念。一个汽车活塞应该是圆的。然而，除非你有办法测量出这个活塞的具体圆度，否则这句话没有任何意义。因此，要改善一个产品的质量，产品的质量就得是可测的。而要测定某产口的性质。就要做这个具体产品的性质做出很好的定义（如上述汽车活塞的例子）。由于所有这些测量就其本质而言都是可变的，因此在生产过程中需要定出这些测量的参数分布。正如 K·皮尔逊通过对数的变化去寻求事物演进的证据，戴明坚持：管理层有责任监控这些测量分布的参数，改变生产过程的基本方面，以改进这些参数。

我第一次见到戴明是在 1970 年的一次统计会议上。他身材高大，表达重要事情时神情严肃，他的外表看上去令人生畏，这在统计学家中非常有名。在讲学后的评论阶段，他很少发表批评意见。只是在会后他才把发言者拉到一边，批评讲学者的缺失。然而，对他的朋友们来说，这种严肃的面孔并不属于戴明，因为我看到的只是他在公开场合的形象。在私下里，他为人亲切，替同事着想，处事稳健，机智、幽默。他热爱音乐，除了参加唱诗班，他还当鼓手，吹奏长笛；甚至还曾发表过几个宗教音乐作品。在他发有的音乐作品中，有一首是为《星条旗》（Star-Spangled Banner）重新谱曲的，据说他说比通常的那个更容易唱。

戴明 1900 年出生于美国爱荷华州（Iowa）的苏城（Sioux），在怀俄明大学（University of Wyoming）读数学专业时，他对工程学有极大的兴趣。后来，他又从科罗拉多大学（University of Colorado)获得数学和物理学硕士学位。在大学期间，他认识了阿格尼丝•贝尔（Agnes Belle），并和她结为夫妻。1927 年，他们迁往康涅狄格州，戴明开始在耶鲁大学攻读物理学博士学位。

戴明第一次为工业企业工作是在位于依利诺斯州（Illinois）西塞罗市（Cicero）的西方电器公司下属的霍桑（Hawthorne）制造厂[38]，他是一边在耶鲁读书，一边趁时期来打工。

[38] “霍桑制造厂”名称的由来是因为著名的“霍桑效应”（Hawthornee effect）。20 世纪 30 年代，霍桑制造厂进行了一项实验。实验者在工厂的管理上采用了两种不同的方法，并进行比较，看结果有什么不同。实验的结果是失败的，原因在于，这两种方法中不论哪一种，参与实验的工人都知道自己是在被仔细观察着的，因而他们的工作效率都得到了明显的改善。从这件事情之后，凡是仅仅因为是做实验就得到改善的事情，就被称为“霍桑效应”。在进行医疗实验时，这种效应会导致实验中的新医疗方法和传统医疗方法两者之间进行比较时发生困难，患者健康状况的改善要比预期的更为明显。

当时在新泽西州贝尔实验室的沃尔特·休哈特已经为统计质量控制方法奠定了基础。西方电器公司作为同一公司（AT&T）下属的一部分，申请在霍桑制造厂实施休哈特的方法。然而，戴明认为，他们并不真正了解休哈特的方法。统计质量管理方法成了基于事先设计好的不允许变动范围的机械程序。而所设的变动范围，往往会使一个不合格品有可能以 5%及以下的机会通过质量控制。之后，戴明以这种质量控制方法会使 5%的消费者不满意为由而将其否定。

1927 年，戴明从耶鲁拿到学位后到美国农业部工作，在 12 年中，主要从事抽样技术和实验设计工作。之后，他离开农业部自己开了家咨询公司，并开始就制造业的质量管理开展培训工作。二战时期培训规模扩大了，当时他培训了近 2000 名设计人员和工程师。这些人回到自己公司后也开办类似的研讨班，到二战结束时，戴明的信徒已达到 30000 人之多。

1993 年 12 月 10 日，最后一次戴明学术研讨会在加利福尼亚（California）举行，戴明以 93 岁的高龄参加了会议，当然研讨会的主要工作都是由他的年轻助手来做的。12 月 20 日，戴明在他华盛顿的家中去世。也就在同一年的 11 月，他的家人和朋友成立了 W·爱德华兹·戴明学会（The W. Edwards Deming Institute），其宗旨就是要促进对戴明的管理思想体系的深刻理解，以推进商业的进步、繁荣和安宁。

戴明与假设检验

在第 11 章，我们提到的 J·奈曼和 E·皮尔逊在统计假设检验方法上所做出的贡献，以及统计假设检验方法是如何在现代统计分析中取得其重要地位的。然而，戴明却对统计的假设检验提出强烈的质疑。他嘲笑假设检验的广泛应用。因为他认为。统计假设检验的研究方向完全聚焦在一个错误的问题上。他直率地指出：“现实当中的问题绝不是两种处理（A 和 B）的差异是否显著。给一个差异，不管它（差异）有多小……我们都会发现……这种（可产生显著性的）实验一直都重复出现。”因此，在戴明看来，仅仅发现显著性差异，没有任何意义，重要的是差异大小程度的确定。此外，戴明还指出，建立在某一实验条件下的差异程度会因条件的变化而不同。因此他认为，标准的统计方法已无法解决其自身的问题。统计学方法上的这些局限性是重要的。戴明指出：“统计学家必须更加关注实际问题，认识和教授统计推论时要看到它的局限性。从一系列结果中越深入地认识到一个推论的局限性，这个推论就变得越有用。”

在本书的最后一章里，我们将会关注戴明在本章中所警告过的统计推断的局限性。

第 25 章 来自黑衣女士的忠告

虽然在 20 世纪初期，统计学方法的发展一直是由男性统计学家占据着主导地位，但是，到了 60 年代，当我步入这一领域时，许多女性占据了重要地位，产业界和政府部门更是如此。例如美国氰胺公司（American Cyanamid Company）的朱迪思·戈德堡（Judith Goldberg）和强生医药公司（Johnson Pharmaceuticals）的保拉·诺伍德（Paula Norwood）都已成为公司统计部门的领导人物。梅维斯·卡罗尔（Mavis Carroll）则是通用食品公司（General Foods）数学和统计服务部的负责人。在华盛顿，女性统计学家担负着人口普查局（Census Bureau）、劳工统计局（the Bureau of Labor Statistics）和国家健康统计中心（the National Center for Health Statistics）等许多部门的工作。在英联邦、在欧洲大陆的其他国家也是如此。在前面的第 19 章里，我们已经看到了她们当中的一些人在推动统计学方法论研究的发展上所起到的作用。

对于在统计史上留名的女性来说，没有谁的经历是典型的，她们都很优秀，她们的个人发展和成就都是独特的，在此，我无法说她们当中的哪一位是女性统计学家的代表，这就如同无法说哪位是男性统计学家的代表一样。但不管怎样，在这里，浏览一位女性统计学家的职业生涯还是很有趣的，这位女士在产业界和政府部门都做出了杰出贡献，刀子就是英国皇家统计学会第一任女会长斯特拉·坎利夫（Stella Cunliffe）。本章中的许多叙述，都是摘自 1976 年 11 月 12 日她在一年一度的统计年会上所做的专题演讲。

凡是认识坎利夫或与她共过事的人都能够体会到她那不同寻常的幽默感，她的机智、敏锐，以及在处理复杂问题时的非凡能力——能够以简单的数学术语解释复杂的数学模型，使她的合作者很快就明白。大量这样的内容出现在她的演讲中——呼吁皇家统计学会的会员们，不要总是停留在抽象的理论研究上，应该多和其他领域的科学家合作。她举例说：“我们经常嘲笑社会学家的分析方法过于粗略，然而作为统计学家，除非我们能为他们提供一些更加科学、更易接受的思想，否则又有什么资格嘲笑他们呢？要实现这一点，我们之间应该是互动的。”她经常举例说明，在实验过程中往往会发生一些事先无法预料的事情。“即使在一个组织完善的实验站进行的大麦试验，也有可能因为拖拉机手的一时鲁莽而前功尽弃——他为抄近路赶回家喝茶而压了实验地块。”

30 年代末期，坎利夫在伦敦经济学院（London School of Economics）学习统计学。在那儿的那段时间是令人激动的。当时，许多学生和一些教授志愿到西班牙去参加反法西斯的西班牙内战，而一些著名的经济学家、数学家及其他学科的科学家，为逃避纳米德国的迫害来到英国，很多人就在坎利夫所就读的学校得到一个暂时的教席。当她完成学业，走入社会时，全世界依然处在大萧条之中。唯一能找到的工作是丹麦的培根公司（the Danish Bacon Company）。她写到：“在那里，用得上数理统计的地方极少，尤其我又是一个女性统计学家，所以在人们的眼里就更加古怪了。”随着二战的来临，她开始参与食品的配给工作，而刀子的数学才能也因此变得有用起来。

在战争结束两年后，她作为志愿者在被战火毁坏的欧洲做救济工作。她是第一批进入荷兰鹿特丹（Roterdam）的人。当时德军正在投降，当地的居民都在忍饥挨饿。在贝尔根－贝尔森（Bergen-Belsen）集中营的受害者被解救出来不久，她就前去给予帮助。她在英国占领区的难民营努力完成了工作。当坎利夫离开志愿者工作时，她已变得身无分文。她找到两份工作，一份是政府食品部下属的油脂部，另一个是英国的吉尼斯本酿造公司（the Guinness Brewing Company），她选择了后一个工作。回想一下前面第 3 章提到过的以“学生”作为笔名发表论文的威廉·S·戈塞特，在坎利夫到吉尼斯之前，他已经在吉尼斯酿造公司建立

了统计部。坎利夫是在他死后 10 年才到吉尼斯公司的，但在吉尼斯，他的影响力仍然很大。人们都很尊敬他，一直都还依据他所创立的实验原则进行科学工作。

统计学在吉尼斯

吉尼斯公司的员工们一直依赖自己的产品。同时为改进自己的产品，一直坚持搞实验。他们

> 从不停止实验，以努力生产水平如一的产品。因为制造啤酒的原料总会受到气候、土壤、啤酒花、大麦不同的影响，还要尽可能地降低成本。人们也许知道也许不知道，由于对自己产品的自负，1929 年之前他们没有做过任何广告。吉尼斯的人认为，吉尼斯的啤酒是能喝到的最好的啤酒，应该靠质量而不是广告去卖酒。至于那些没喝吉尼斯啤酒的人，只能为他们感到惋惜，而不是向他们打广告！直到我离开公司时，他们还是这个特性。

坎利夫描述了她第一天来到吉尼斯时的情景：

> 到都柏林来“实习”的生活，就如同在德国时一样自由而又充满刺激。一天早晨，当我出现在都柏林酿造公司专管女职工的女主管面前时，只见她一身黑衣，领口镶着一圈花边，用鲸骨撑着……，表情严肃，她告诉我，能被选来吉尼斯工作是一种殊荣，并提醒我应该穿长筒袜，戴帽子。如果在走廊里有幸碰到某一位“酿造师”——公司的重要人物，不管认识与否，都要低头为他让路。

这就是 1946 年，妇女在等级分明的吉尼斯公司的地位。

坎利夫很快就证实了自己对公司的价值，并深入到公司在爱尔兰的农业实验中。她不喜欢一天到晚坐在办公桌前分析野外科学家采集来的现成数据，而是到野外实验基地去，亲自了解实验的动态。（任何一个新任的统计工作者都应该以她为榜样。一个令人惊异的事实是，那些比实验室普通员工高出好几级的高级管理者们，他们所做出的实验结论，往往是与实际不符的。）

> 不知有多少个阴冷潮湿的清晨，刚刚 7 点钟，我便来到啤酒花实验园。虽然又冷又饿，但那是在参与“至关重要”的实验。我之所以要用“至关重要”这个词，就是因为如果统计学家本身都不重视这样的实验，那么又怎能调动起实验参与者的激情，使他们做出最佳贡献呢？但是，作为一个统计学家，我们必须学会灵活机动，要能适应转辗于各种不同类型的工作。或许我们得帮助一个生物学家进行新酵母菌的实验；去帮助一个农业专家完成另一项实验——了解以一种特殊饲料喂养的家畜的粪便变化情况；与病毒学家讨论为纽卡斯尔（Newcastle）病毒研制的新抗体；去协助一个医疗官员评估麦芽储藏中的灰尘对人体健康的影响；去给一个正在进行传送带实验的工程师提些建议；试着将统计的排除论（queuing theory）应用到职工餐厅的管理上；或者去协助一个社会学家验证他的群体行为理论。

以下是一份为产业界工作的统计学家的典型协作类型清单。根据我本人的工作经验，我们要与之打交道的人包括化学家、药理学家、毒理学家、经济学家、临床医生、经营管理者（我们为他们开发运筹模型用于决策）。可以说，数理统计方法的应用无处不在，作为数学模型专家，统计学家可以与任何领域的人合作，为他们提供服务。这也许就是统计学家的工作之所以迷人的原因之一。

非预期的变异

在坎利夫的演讲中，她指出，最大的变异还是来自现代人类本身。

在吉斯尼期间，我很高兴负责组织对啤酒的品尝实验，对于吉尼斯啤酒这一美好饮品的发展来说，这无疑是一项非常有意义的工作。通过这些实验，我开始认识到，人在不可能没有偏好，没有偏见，没有最感兴趣的事，但这也正是让人着迷的地方。我们都有喜好的某些数字、字母、颜色，实际上，我们都是特别迷信的。我们都有非理性的行为。在我的记忆中，曾开展过一次大型的有关啤酒温度的实验。让一些人在不同室温的环境中品尝不同温度的啤酒，以判断人们对不同温度啤酒的喜好程度。当时，一些身着白色制服的人跑上跑下地送酒，啤酒则放在不同温度的水桶中，每个桶里都配有温度计。实验场所一片喧嚣，啤酒用不同颜色的瓶盖来辨认。最后实验惟一明确的结果是，受试都只在意瓶盖的颜色，他们不喜欢黄颜色瓶盖的啤酒，至于啤酒的温度，几乎没人注意！

坎利夫还讲述了一个检验小号啤酒桶容量的事。这些桶是手工制造的。需要检验它们的容量以确定其尺寸是否符合规格。进行检验的女工首先称一下空桶的重量，然后再将桶中灌满水，称一下装满水的桶的重量。如果桶的重量比标准的少 3 品脱以上，或多 7 品脱以上，则都作为不合格品退回去返工。作为进行质量管理程序中的一部分，统计人员负责检查合格品与不合格品的报表情况。在检查满桶重量的记录图时，坎利夫发现，刚刚在合格线以里的桶数目过高，而刚刚在合格线外的桶数目过低。为此她到现场了解工作条件，发现女工必须将挑出来的不合格品堆放到旁边的一大堆桶上，而合格品只需放在传送带上即可。于是她建议把女工的座位加高，工作时只需将挑出来的不合格品直接踢到脚下的箱子里即可。结果发现，报表上反映的合格率，很快趋于正常水平。

坎利夫后来成为吉斯尼统计部的主管。1970 年，她被调到英国内务部（the British Home Office）调查局，这个单位负责警察、法院和监狱的监督工作。

在我刚来的时候，这个单位主要从事与犯罪有关的工作。说老实话，我在吉斯尼公司所从事的是十分精确、精心设计、透彻分析的统计工作，而这里要做的都是些社会学家，有时是心理学家的工作，在我看来是一个空泛的世界。我丝毫没有贬低内务部调查局研究人员能力的意思。但是无论如何，使我震惊的是不同的工作原则：设立一个零假设，制定细致详尽的实验设计方案，抽取足够的样本数量，进行小心谨慎的统计分析，做出详细的评估结果，所有这些都是我长期做过的，到了社会学这个领域，却都变得微不足道。

在刑事犯罪学这一领域的主要研究工作，就是积累长期的数据资料，进行分析，以发现公共政策对其可能的影响。如有一项分析是针对男性囚犯进行的，即研究不同刑期的男性囚犯出狱后两年内重新犯罪的概率有多大。分析结果清楚地表明，刑期越短，重新犯罪的概率越高。从而作为一个证明：长刑期可以把惯犯从街头清除。

坎利夫并不满足于重犯率与刑期间简单对比的数表。她要进一步分析数表背后所隐藏的东西。这种明显强关联关系主要是刑期在三个月之内的犯人重犯率高，经过仔细检查，这些人“几乎都是些年老的、处境悲惨的、精神不正常的人，他们被精神病医院拒之门外，所以才一次又一次地反复犯罪后再进监狱。”而统计表所反映的数字，实际都是这同一拨人，被当作不同人重复统计，才将短期犯罪的重犯率夸大了。统计表中的另一个极端表现是，刑期在 10 年以上的犯人出狱后只有 15%的人又重新犯罪。坎利夫认为，“这里有一个很大的年龄因素，一个很大的环境因素和一个很大的犯罪程度因素。刑期长的都是些犯大案的人，他们出狱后重新犯同样大案的可能性也不大了。”因此，在她用两个极端的情况将数表调整这后，重犯率和刑期间明显的关系消失了。

坎利夫说道：

我认为，即使所谓单调的内务部统计仍然是很迷人的，……对于我来说，统计学家的工作就是阅读数据，并质疑它们为什么会是这样的？……我今天晚上来此的想法很

简单，就是想告诉人们，数字是很有趣的，如果听众当中有人感到枯燥，那是我们没有表述好，或者因为数字本身的问题。不过，根据我在内务部的统计工作，我要说，数字一点也不枯燥。

坎利夫谴责政府官员中的一种不良倾向，他们在决策时没有仔细研究阅读手头的数据资料。

> 这不是社会学家、社会工作者和计划制定者们的过错，但却是统计工作者不可推卸的责任。我们还没有学会去为这些在我们看来不那么科学的学科服务，因此我们还没有作为能帮助他们增进知识的人而被接受，……根据我的经验，统计学家在应用领域的力量……在于他（或她）说服他人的能力：去形容所需回答问题；去考虑实验员可用的工具是否足以回答这些问题；去帮助他建立合适的零假设；去实施严格的实验设计原则。

据我本人的经验，将问题尽可能地以数学模型的方式表述出来，这会迫使科学家去充分了解将会产生什么样的问题。仔细地检查可利用资源，经常会得出这样的结论：用这些资源是不可能回答出该问题的。我想，作为统计学家，我的主要贡献之一，就是阻止别人去尝试因缺乏适当资源而注定要失败的实验。例如，一项临床实验，其中的医学问题需要有数十万名病人的配合。这就有必要考虑这个问题是否值得回答。

抽象的数学还是实用的统计学

坎利夫特别看中那些对统计分析有用的预期工作，她轻视为数学而数学的推敲，她诋毁下面的那种数学模型：

> 全是空想，缺乏实际，很多线索，有趣的片断，充满趣味性，精彩的概念，但同时也缺乏稳健性。这种高雅的乐事往往是以牺牲实践性为代价的，恕我直言，在我看来它似乎更合乎男性的口味。我们统计学家所受的教育就是进行计算，同时要考虑数学的精确度。我们并不善于说服那些毫无经验的人，让他们知道我们的发现值得注意。如果我们一本正经地对一个不懂统计的男人或女人说“P 值小于 0.001”意味着什么，我们就不会成功，所以，我们必须用他们的语言来解释我们的发现，以增强说服工作的效力。

不戴帽子，不肯向酿酒师这样的“大人物”低头，坎利夫进入了统计的世界，她尽情地满足了自己强烈的好奇心，她批评那些来听自己演讲的数理统计学教授。当我写这本书时，她可能仍在皇家统计学会用她那辛辣的机智，表现她的数学主张。

第 26 章 鞅的发展

充血性心脏衰竭是世界上致人死亡的重要原因之一。虽然这种病在壮年人当中也时有发生，但此病主要还是一种老年性疾病。以美国为例，在 65 岁以上的老年人当中，有半数是死于充血性心脏衰竭或它的并发症。从公共健康的角度来看，充血性心脏衰竭不仅是致人死亡人重要原因，也是引发生活中诸多其它疾病的一个重要因素。此外，患者为稳定病情而反复住院，以及治疗过程的复杂程序，是导致国家的公共医疗服务成本居高不下的一个重要因素。为此，许多人都殷切希望能找出更好的辩论治疗方法，以减少患者住院治疗的需求，同时改善这些病人的生活质量。

不幸的是，充血性心脏衰竭不是一种普通的疾病。其病因不是一种简单的传染源，也不能通过阻断某种生化酶的通路而缓解。人体中荷尔蒙精巧地控制着心脏，调节其跳动的速度和收缩能力，以适应身体变化着的需求，但充血性心脏衰竭患者的心脏对这种调节的反应能力越来越差，患者的主要症状表现为心肌逐渐衰弱，心脏的肌肉变得越来越肥大、松弛。患者会因此而出现肺部和脚踝的水肿，轻微的运动都会导致他们呼吸困难。患者还会因进餐时胃部供血而造成的脑部供血不足而感到困倦和意识混乱。

为保持体内平衡，病从的身体会自动调节以适应心脏能量输出的减少。对许多患者，调节心肌和其它肌肉变化的荷尔蒙会在某种稳定状态达到平衡。虽然就一般人来说，这样的荷尔蒙水平是不正常的。如果医生在治疗过程中使用了 β 肾上腺素收缩剂或钙离子隔断剂，结果可能使患者的情况变得更为复杂。肺部水肿是充血性心脏衰竭病人死亡人一个重要原因。现代医学依靠利尿剂这种药物可以使水肿得到缓解。然而，患者在使用了利尿剂后，为调节肾功能和心脏功能所导致的荷尔蒙的变化，又会因相互影响而造成新的难题。

长期以来，医学界一直致力于研究更加有效的治疗充血性心脏衰竭的方法，希望延长患者的生命，减少他们的住院次数，提高他们的生存质量。由于一些治疗可能会对某些病人产生不良影响，因此，治疗的任何可能会对某些病人产生不良影响，因此，治疗的任何临床研究都需要考虑到特殊病人的情况。在这种情况下，这种研究的最终数据分析可以指认对哪些病人有效，哪些病人有不良反应。所以，对充血性以及衰竭研究的统计分析将变得难度极大。

当设计一项研究时，首先遇到的问题是要测量什么。例如，测量某一种治疗患者的平均住院治疗时间，这是一种粗略的总体测量，没有考虑到重要的方面，如他们的年龄，他们最初的健康状况，他们发病的次数，以及住院治疗的时间。最好要考虑到每个患者发病的整个时间过程，估量可能的住院治疗情况，如住院时间长度，与上一次住院治疗的间隔，出院期间患者的生活质量，并根据患者的年龄以及其它可能发生的疾病，对所有这些结果进行调整。从医学的观点来看，这可能是一个理想的方案，但它提出了一个困难的统计学问题。这里没有一个数据与单个患者相联，相反，患者的记录是事件的时间过程，有些记录是重复的，有些通过多重测量得到。因为在这个试验中的测量是多层次的，因此，其分布函数——这些函数的参数必须是可估计的，其构成也必须是多维的。

早期的理论性工作

解答这个问题，是从法国的数学家保罗・利维开始的。保罗・利维出身于数学世家，他的父亲、祖父都是数学家。保罗・利维 1886 年出生，在他还很小的时候，就显示出与众不同的学习天赋。按当时漫无边际的惯例，他很快升入专门培养天才学生的学校，并且在学

习期间获得过许多学术性奖励。还是十几岁少年的他，就获得了希腊文和数学的法国中学中学优等生会考奖；获得法国国立圣路易学校（Lycée Saint Louis）颁发的数学、物理学和化学的成绩优异奖；获得了高等师范学院及综合工科大学入学竞赛第一的成绩。1912 年，26 岁的保罗·利维获得科学博士学位，他后来写的一本有关抽象函数的重要著作，就是以他的博士论文为基础的。保罗·利维获得科学博士学位，他后来写的一本有关抽象函数的重要著作，就是以他的博士论文为基础的。保罗·利维在 33 岁时就成为综合工科大学的全职终身教授，法国科学院院士。他在抽象分析理论方面的工作使他闻名于世。1919 年，他所在学校安排他就概率论问题开展一个系列讨论，为此，他首次着手就这一问题展开深入的研究。

利维不满于当时作为复杂计算方法之集合的概率理论（那里安德烈·柯尔莫哥洛夫的理论尚未出现）。利维寻找一些基础性的抽象数学概念，以便把这些方法统一起来。在这一过程中，棣莫弗正态分布的推导和数学家的“大众定理”（FOLK theorem）打动了他。（按大众定理，棣莫弗的结果在许多其它情况下也都成立，现在叫做“中心极限定理”）我们已经看到利维（与荷兰的林德伯格（Lindeberg））如何在 20 世纪 30 年代早期最终证明了中心极限定理，以及这个定理成立的必要条件。与此同时，利维着手对正态分布公式进行研究，通过逆向推导，寻求这一分布的独特性质，使得该分布能由这么多的情形产生出来。

然后，利维又另辟新路，从另一个角度探讨这个问题，探询这种正态分布成立的特定条件是什么。他确定只需两个简单的条件就能使一组数列趋向一个正态分布。但这两个条件并不是正态分布能产生的唯一途径，利维对中心极限定理的证明建立了一组更具有普遍意义的必要条件，这两个条件相当于有一组随机产生的一个接一个的数列：

1. 变异是有界的，因此个别值不可能是无穷大的，也不可能是无穷小的。
2. 下一个数字的最佳估计值必是它的前一个数值。

利维称这样的数列为鞅（martingale）。

这里，利维借用赌博中的一个术语。在赌博中，martingale 的意思是指赌博者在输了的情况下加倍下注，如果他输赢的机会各半，即 50%：50%，那么损失的期望值就等于他原来的损失。Martingale 这个英文词还有另外两个含义。一个意思是用来描述法国农夫套马的一种装置，让马低着头不向后甩。在此装置控制下，马的头可以随意活动，但马头下一个最有可能的位置是它现在所在的位置。Martingale 的另一种解释是用在航海上的。指一片很重的木头，悬挂在船帆的下桁上，用以防止帆的下桁因剧烈摇晃而左右摆动。这里，帆的下桁最后的位置也就是它下一次位置的最佳估计。至于这个词本身，是来源于法国的一个叫马提克（Martique）的小镇，该小镇的居民以小气而著称。据说他们下周要花的一点小钱，估计起来最有可能等于他们今天花的钱。

利维正是从马提克小镇居民的小气习性中受到启发，创立了最小气可能性的抽象数学概念，而具有这种性质的数列通常是正态分布的。到 1940 年，鞅已经成为抽象数学理论的一个重要的工具。它的简单必要条件，意味着诸多类型的随机数列都具备鞅的性质。1970 年，挪威奥斯陆大学（the University of Oslo）的奥德·奥伦（Odd Aalen）研究发现，在临床试验中，病人的反应方式就是一个鞅。

鞅与充血性心脏病研究

回想前面有关充血心脏病研究所引发的问题，因为患者的反应各不相同，我们的问题就是如何解释研究中患者住院治疗的时间早晚问题（当患者年龄已经很大的时候），如何处理患者住院治疗的次数和住院时间的长短。把长时间得到的数据看成鞅，所有这些问题的答案都可能回答。奥伦特别注意到，当一个患者住院治疗时他可以从分析中排除，到其出院后再列入研究范围。重复多次的住院治疗可以把每次住院作为一个新事件来处理。在每一个时

间点，分析人员需要了解的就是仍在研究中（或回到研究中）的病人数和最初进入研究的病人数。

在 20 世纪 80 年代初，奥伦与丹麦奥尔胡斯大学（Aarhus University）的埃里克・安德森（Erik Anderson）及荷兰乌得勒支大学（University of Utrecht）的理查德・吉尔（Richard Gill）一起探索他的新发现。在本书的第 1 章我就曾指出，数学的发展总是和科学发展具有不可分割的联系。抽象的数理统计是如此错综复杂以至于很容易出错，只有通过同事间共同的讨论和批评，才能发现其中可能出现的错误。正是奥伦和安德森、吉尔这三个人的通力合作，造就了 20 世纪最后十年这个领域的一项最富成效的研究结果。

之后，理查德・奥尔森（Richard Olshen）与其在华盛顿大学的合作者，以及哈佛大学的魏立人（Lee-Jen Wei）教授又对奥伦、安德森和吉尔三个人的研究成果进行了补充。他们又提出了大量用于分析临床试验中序列事件的新方法。特别是魏立人对于两个鞅之差仍然是鞅这一概念的开拓性的应用，消除了对模型进行多个参数估计的必要性。如今鞅方法在慢性疾病的临床试验研究统计分析中占据着主导地位。

以马提克上镇居民以小气著称的传奇故事为起点，法国人利维创立了建立在最小气原理的数学概念之上的鞅方法的最初概念。之后，又经过更多头脑的共同研究开发，他们包括美国人、德国人、俄国人、英国人、意大利人和印度人。之后，又由挪威人、丹麦人和荷兰人将这种方法运用于临床试验研究。两个美国人，其中一个出生在中国的台湾，又进一步将这项研究推向深入。20 世纪 80 年代以来，有关这方面问题研究的文章和书籍特别多，光是作者名录就可以写好多页，研究者还来自上面没有提到的很多国家。的确，数理统计学已成为一种国际合作性的研究。

第 27 章 意向治疗法

在 20 世纪 80 年代初，英国杰出的生物统计学家雷沙尔·皮托（Rechard Peto）遇到了一个难题，当时他正在分析比较不同癌症治疗方法的临床试验结果。根据费歇尔实验设计规定，典型临床实验研究要求确定需要治疗的病人群体，并且采用随机的方法分配给病人不同的治疗实验方法。

数据的分析应该是相当直接的，用费歇尔方法，只要在不同治疗方法的组别间，比较病人的 5 年存活率即可。另外还可以进行更加精确的比较，就是用奥伦（Aalen）的鞅方法（martingale approach），分析从开始研究到每个病人死亡的时间，以此作为衡量治疗效果的基本标准。不论是哪种方法，分析结果的准确性取决于最初分配给病人采用治疗方法的随机选择。根据费歇尔定律，指定病人采取何种治疗方法与研究的结果是完全不相关的，假设检验的 P 值是可以计算出来的。

皮托的难题是所有病人的治疗方法并不是随机指定的。这些病人也是人，正饱尝病痛的折磨，而且很多人得的是绝症，因此医生沉得有责任放弃实验性的治疗，或者如果觉得对于病人来讲是最好的选择的话，至少也要进行方案的调整。盲目地照搬某种治疗方法而不考虑病人的需要和反应是不首先的。与费歇尔的实验设计要求相矛盾，在这些实验中的病人经常变换治疗方法，而对治疗方法的选择主要取决于病人的治疗效果，如果效果好可能会继续采用这种方法，一旦觉得治疗效果不理想就会改变治疗方法。

这是癌症研究中的一个典型问题。从 20 世纪 50 年代人们刚刚开始研究癌症起，这就一直是一个令人困扰的问题，直到皮托涉入此领域研究之前，通常的做法只是去分析那些坚持采用随机分配治疗方法的病人，而其他的病人不在分析的范围之内。皮托认为这会导致严重的错误。例如，假设我们正在比较两种治疗方法，一种是有效的治疗，另一种只是给病人服用安慰剂，即一种没有生物作用的药物。如果病人对治疗无反应，就会转而使用常规的治疗。服用安慰剂、没有效果就转而使用别的治疗方法的病人不能做为研究对象，只有那些继续服用安慰剂、因为某些原因有反应的病人才是研究的对象。如果在研究分析中的研究对象只有那些继续服用安慰剂并且有反应的病人，那么研究的结果必然是：安慰剂治疗方法与有效的治疗具有同样的疗效，甚至可能疗效更好。

德克萨斯州安德森医院（M. C. Anderson Hospital）的埃德蒙·吉亨（Edmund Gehan）比皮托更早发现了这个问题。他当时的办法只是提出：因为这些研究不符合费歇尔实验的条件，所以不能够作为比较不同治疗方法的有效实验，只能算是研究中通过对采用不同治疗方法病人仔细观察而取得的记录，最多只是对实验结果的一种总体描述，为以后的治疗提供了一些思路。后来，吉亨也考虑了解决这个问题的不同方法，但是他的第一个结论让人非常气馁，竭力想在一个设计和执行都不好的实验中运用统计分析方法看来是不可能的。

皮托提出了一个直截了当的解决方法：当比较不同的治疗方法的疗效时，病人采用哪种治疗方法应该是随机的，否则不可能在假设检验中计算出 P 值。他建议在分析过程中假定每个接受治疗的病人采用治疗方法是随机分配的，否则不可能在假设检验中计算出 P 值。他建议在分析过程中假定每个接受治疗的病人采用治疗方法是随机分配的，忽略研究中治疗方法的调整。如果一个病人随机采用方法 A，但在研究结束前改变了方法，这个病人视为采用 A 方法的病人进行研究；如果病人随机采用方法 A 只治疗了一个星期，病人当作采用方法 A 来分析；如果病人随机采用 A 方法治疗，却根本没有吃一粒 A 方法的药，就采用了另外一种治疗方法，这个病人仍被视为采用方法 A 的病人。

乍一看这种方法是愚蠢的。人们可以假设一种情形：对一个实验治疗方法和一个标准治疗方法进行比较，病人采用的实验治疗方法一旦失败就会转而使用标准方法。如果实验治疗方法是无用的，那么，所有的或者大多数被随机指定使用实验治疗方法的病人就会转而使

用标准方法，分析将会发现这两种治疗方法效果是一样的。正如皮托在他的假设中指出的，这种分析研究结果的方法不能用于比较疗效相同的治疗方法，只有当疗效“不同”时才可使用。

皮托的方法后来被称为“意向治疗”（intert to treat）分析方法。这样命名的理由及其用途是：如果我们对医疗政策的总体结果感兴趣的话（该政策通常会推荐使用某个治疗方案），就得授权引而伸之医生，让他可以按照他的判断去调整治疗方法。用皮托的方法，临床实验的分析可以判断：建议使用一个给定的方法作为治疗的起点，是不是一个好的公共政策。“意向治疗”分析方法最被认为是一种很好的方法，适合用于那些政府资助的、为制定好的公共政策而进行的大型研究。

很不幸的是，有些科学家往往在并不了解和理解其背后数学含义的情况下，随意地把一些统计方法拿过来就用，这在临床研究中是司空见惯的。皮托早就指出了他的方法的局限性，但是意向治疗方法不但已经成为许多大学里的医科教条，并且被认为是临床实验唯一正确的统计分析方法。在许多临床实验中，尤其是对癌症的研究实验，实验设计是为了证明新的治疗方法至少与标准治疗方法效果相同，同时副作用较小。很多的实验目的是为了显示新疗法的等效性。正如皮托指出的，他的方法只能用来找出差别，但是，如果没有找出差异也并不代表两种方法的疗效相同。

某种程度上，这个问题的产生主要是因为奈曼－皮尔逊理论的刚性。在基础统计学的教科书里都可找到奈曼－皮尔逊理论的标准版本，假设检验往往被介绍为一种固定的程序，方法中许多完全随意的方面也被描述成不变的。

尽管许多这些随意的元素并不适用于临床研究[39]，但是一些医学家在研究中不得不用“正确”的方法，这种需求使得他们视奈曼－皮尔逊理论为最严格的信条，除非通过统计程序事先确定了 P 值，并且使之保持不变，否则没有任何事是可接受的。这是费歇尔反对奈曼－皮尔逊理论的原因之一，他认为 P 值和显著性检验的应用程序不应该受如此严格条件的限制，他特别反对奈曼事先竟然确定了错误概率的存在，并且只有在 P 值小于这个事先确定的值时才有效。费歇尔在《统计方法和科学推论》（Statistical Methods and Scientific Inference）一书中建议，对于 P 值多大才有意义，最后结果应视情况而定。在这里我用了“建议”的字眼，是因为费歇尔从没有很明确地说明他怎么使用 P 值，他只是提供一些例子。

考克斯的理论

1977 年，大卫・R・考克斯（即第 23 章里提到的博克斯和考克斯中的一位）开始研究费歇尔的论点，并对它们加以发展。为了区分费歇尔所用的 P 值和奈曼－皮尔逊理论，他称费歇尔的方法为“显著性检验”（significance testing），而称奈曼－皮尔逊的理论为“假设检验”（hypothesis testing）。在考克斯撰写他的论文的时候，统计显著性（通过计算 P 值）的计算已经是应用最广泛的科学研究方法，因此，考克斯断言，这种方法已经证明了其在科学研究中的作用，尽管存在费歇尔与奈曼之间的尖锐争执，尽管存在 W・爱德华兹・戴明这样的统计学家坚持认为假设检验毫无用途，尽管出现了根本不需要计算 P 值、不需要考虑显著性的贝叶斯统计学……总之，尽管在数理统计学家之间存在着上述这些争论，显著性检验和 P 值一直被使用着。考克斯就问了：科学家真的在使用这些检验吗？他们怎么会知

[39] 1963 年，耶鲁大学的弗朗西斯・安斯孔（Francis Anscombe）提出了一种完全不同的方法，这种方法更符合医学的需要。奈曼－皮尔逊理论使分析人员发生错误的概率增大。安斯孔要问的是：为什么统计分析人员的长期误差概率会与决定一个治疗方案的是否有效相关联？作为替代，安斯孔提出，接受治疗的病人人数是有限的，他们中只有少部分人接受临床实验治疗，其余的病人将采用临床中认为是“最佳”的治疗方案。如果我们在实验中的病人过少，那么在确定哪一种治疗方案时主会产生误差，并且，如果这样，其余的病人将会采用错误的治疗方法。如果我们在实验中采用了太多的病人，那么采用另外治疗方案（不是“最佳”的）的病人将会被安排采用错误的治疗方法。安斯孔提出，分析的准则应该是将采用了较差疗法的病人（包括临床实验病人和后来接受治疗的病人）数量降到最少。

道这些检验的结果是真的还是有用的呢？他发现，在实践中，科学家用假设检验主要是通过消除不必要的参数，来提高其对现实的了解程度，或是用来在两个不同的现实模型间进行选择。

博克斯的研究方法

博克斯（博克斯和考克斯中的另一位）从稍微不同的角度来研究这个问题。他认为，科学研究不只是做一个简单的实验，科学家在进行实验前，已经掌握了大量的知识，或者至少对实验的结果已经有了一个期望值，研究是为了提升知识、实验设计取决于你要提升的知识类型。在这一点上，博克斯和考克斯具有很多共同之处。对于博克斯来说，一次实验是一系列实验的一部分，将这次的实验数据与其它实验的数据进行比较，那么早先的知识就会在新的实验中和对以往实验的重新分析中得到重新审视。科学家从未停止过对以往研究的回顾，并从较新的研究视角去提升过去的认识。

举一个关于博克斯方法的例子。假设一个造纸厂引进了博克斯的一个主要创新方法——调优运算（evolutionary variation in operations，EVOP），按照博克斯的方法，这个工厂在生产过程中引入了一系列的实验，用不同的方法在温度控制、速度、硫磺处理过程以及温度控制等环节进行了微调，结果发现纸张的强度变化不大。如果要生产的产品仍然可销售的话，这种变化是不能大的。然而，根据费歇尔的方差分析（analysis of variance），用这些微弱的差别可以进行另外一个实验，在这个新的实验中，纸的平均强度稍微增大，这样，这个新的实验就可以用来确定可以提高纸张强度的工作方向。在过程操作改进中每个步骤的结果都与先前步骤的结果进行比较，当得到的结果看起来比较反常时，实验要重新做，这个过程周而复始——永远没有所谓最终“正确”的结论。在博克斯的模型里，这个不断进行着数据检验和再检验的科学实验是没有尽头的——没有最后的科学真相。

戴明的观点

戴明和其他许多统计学家坚决否定假设检验的作用。他们坚持认为费歇尔的估计方法才是统计分析的基础，认为真正应该估计的是统计分布的参数，而通过 P 值和武断的假设间接地处理这些参数而进行的分析是毫无意义。这些统计学家继续使用奈曼的置信区间去衡量他们研究结论的不确定性，但是他们却认为奈曼－皮尔逊的假设检验就象 K·皮尔逊的矩法（method of moments）一样已经过时了。有趣的是，奈曼自己也很少在他的应用性论文里用到 P 值与假设检验。

对假设检验的拒绝以及博克斯与考克斯对费歇尔显著性检验定义的重新诠释，使得人们可能对于皮托在癌症临床研究中解决问题的方法提出质疑。但是他面对的这个根本问题始终没有解决。当接受治疗的病人改变治疗方法，实验因此被动地做了调整时你能怎么做？亚伯拉罕·沃尔德（Abraham Wald）已经指出在实验中怎样的调整是可以接受的，那就是序贯分析（sequential analysis）。但是在皮托的问题中，肿瘤学家不会采用沃尔德的序贯分析法，一旦他们察觉到必要时，他们就会采用不同的治疗方法。

科克伦的观测研究

从某种方面来说，皮托的问题也是约翰·霍普金斯大学的威廉·科克伦在 20 世纪 60 年代研究的问题。巴尔地摩（Baltimore）市政府想知道，公共住宅是否影响低收入人群的社会态度和生活水平的提高。他们联系了约翰·霍普金斯大学的统计小组，请求他们帮助设计一个实验。按照费歇尔的方法，约翰·霍普金斯大学的统计学家建议寻找一群人，不论他们是否申请了公共住宅，随机分配公共住宅给其中一部分人，而对其中的另外一些人不提供公共住宅。这个建议吓坏了市政官员，以往，在公布安置公共住宅时，他们通常的做法是先到

先受理，这是惟一公平的做法，他们不能拒绝那些先提出申请而却是因为计算机的随机抽取而没有选中的人。但是约翰 • 霍普金斯大学的统计学家指出，不管使用何种方法，那些最先申请的人通常都是最积极并且有野心的人，如果这种说法是对的，那么住在公共住宅里的人本来就比另外一些人干得好，这与提供住宅本身无关。

科克伦的结论是，如果他们不能够采用已经设计好的科学实验，那么通过追踪那些住进公共住宅以及那些没有住进的家庭，他们可以采用观察研究的方法来替代。这些家庭有很多因素不同，如年龄、受教育程度、宗教信仰以及家庭的稳定状况。他对这类观察研究的统计分析提出了许多方法，在各种方法中，他会考虑不同家庭的上述因素对测量结果进行调整，建立一个数学模型，其中包括年龄、是否是单亲家庭、宗教信仰等因素的影响力。一旦代表这些因素的影响力参数估计出来了，剩下的影响就应该是由公共住宅造成的。

如果临床研究声称，治疗效果的差异已经根据病人年龄和性别的差异进行了调整，那就是说研究人员在估计治疗方法的主要效果时，已经应用了科克伦的方法，并且考虑了在治疗中为病人指定方法不平衡性的影响。几乎所有社会学研究都采用了科克伦的方法，但有些研究的作者可能没有认识到他们用的方法来自科克伦，而且认为其中很多特殊技术通常比科克伦的研究还要早。然而，科克伦为这些方法建立了稳定的理论基础，他写的关于观察研究的论文已经影响了医学、社会学、政治科学和天文不，在这些领域里“治疗方法”的随机指派，既不可能，也不道德。

鲁宾模型

在 20 世纪 80 年代和 90 年代，哈佛大学的唐纳德 • 鲁宾（Donald Rubin）提出了不同的方法，来解决皮托的问题。在鲁宾的模型中，假设每个病人对每个治疗方法都有一个可能的反应，也就是说，如果有两个治疗方法 A 和 B，我们可以只观察采用其中一种治疗方法的病人，这些病人采用的方法是已经确定的。我们可以建立一个数学模型，在这个模型的公式中用一个符号来表示每种病人可能会有的反应。鲁宾界定了这个数学模型的使用条件，而在估计病人转而使用其它治疗方法会有什么样的反应时，这些条件是必需的。

鲁宾模型和科克伦的方法可以应用于现代统计分析中，因为应用计算机可以处理大量的数据。这些方法即使在费歇尔时代有人想到了，也是不可能实现的，因为这个数学模型涉及的数据太多，计算非常复杂，必须要借助于计算机。这个方法经常要求进行迭代计算，计算机要进行上万甚至百万次的计算，最后才会收敛于一个最终的答案。

科克伦和鲁宾的方法是高度依赖特定模型的，也就是说，除非所用的这个复杂的数学模型能非常准确地描述现实，否则就不会得出正确的答案。如果使用他们的方法，就要求分析人员要建立一个能够全面或近似全面描述事实各个方面的数学模型，如果事实与模型不符，那么分析的结论就不成立。像科克伦和鲁宾这些方法的一个伴生部分，已经成为去确定事实与模型怎样的拟合度下，结论是稳健的一种尝试。目前，数学界正在致力于研究：在结论不再成立之前，事实与模型之间可以有多大偏差。科克伦在直到 1980 年去世以前的日子里，一直在研究这些问题。

统计分析方法可以看作是一个连续过程，一端是高度依赖模型的方法，如科克伦和鲁宾的方法；另外一端则是一些非参数方法，采用最普通的方式检查数据。正如计算机的出现使模型模拟的方法得以实现一样，在使用非参数方法时，也发起了一场计算机革命，这种方法极少或根本不用设计数学结构，数据不必放在一个预想的模型中就可以展现它们的含义。这些方法在使用中都有一些奇怪的名字，像“解靴带”（“boot-strap”，我们称为“自助法”——译者注）。这是下一章要叙述的内容。

第 28 章 电脑随心所欲

圭多·卡斯泰尔诺沃（Guido Castelnuovo）出生于显赫的意大利犹太家庭，他的家庭背景可以追溯到古罗马最早的凯撒时代。1915 年，卡斯泰尔诺沃当时是罗马大学（University of Rome）的数学教授，他正在进行一场孤独的战争，他想在研究生项目中引入一些有关概率和精算数学的课程。当时，安德烈·柯尔莫哥洛夫还没有建立起概率论的基础，数学家认为概率只是一个使用了复杂计算技术的众多方法的集合，是数学中的一个有趣的花絮，经常作为代数课里的一个部分来教授，在纯数学美丽的微光尚待关注的时候，没有人认为值得在研究生项目中开设这种课程。就精算数学而言，这段时间是应用数学最低迷的时期，人的寿命及意外事故发生频率的计算都只是采用简单算术，所以，系里其他的数学教授都认为没有开设这个课程的必要。

卡斯泰尔诺沃不仅在代数几何学这个抽象领域做了许多开创性工作，他对数学应用也有着浓厚的兴趣，他还劝说系里的其他人允许他开设这个课程。作为教学的成果，他在 1919 年出版了第一本关于概率与统计应用的教科书《概率运算与应用》（*Calcolo della probabilità e applicazioni*），这本书被意大利其它一些大学用于类似课程的教学中。到了 1927 年，卡斯泰尔诺沃已经在罗马大学成立了统计与精算科学学院（The School of Statistics and Actuarial Sciences），而且在整个 20 年代和 30 年代，意大利学校里致力于精算研究的统计学家越来越多，他们与瑞典该领域的专家进行极其活跃的交流。

1922 年，贝尼托·墨索里尼（Benito Mussolini）在意大利实行法西斯主久，利用强权控制人民的言论自由，对大学里的学生和教职工都进行调查，以驱逐所谓的“国家的敌人”。在这次驱逐行动中，因为没有提及种族问题，所以卡斯泰尔诺沃是犹太人这件事没有被考虑进去[40]。所以最初的 7 年里他能够继续在法西斯政府的统计下工作。到了 1935 年，意大利法西斯与德国纳粹的联合导致在意大利实行反犹太的法律，70 岁的卡斯泰尔诺沃失去了工作。

但是，这些并没有使这位不知疲倦的人停止工作，直到 1952 年去世。随着纳粹种族政策的实施，许多有前途的犹太研究生也被逐出大学。卡斯泰尔诺沃就在他和其他犹太教授的家里设立了特殊的课堂，坚持授课，以帮助这些犹太研究生继续他们的学业。卡斯泰尔诺沃除了写一些关于数学历史的书外，还在他 87 岁时的最后日子里，研究决定论和机遇之间的哲学关系，并试图去说明因果的概念——这些我们已经在前面的章节中接触过了，在本书的最后一个章节我将作进一步的探讨。

由于卡斯泰尔诺沃的努力而建立起来的意大利统计学派，拥有稳定的数学基础，但大多数研究都是以在实际应用中遇到的困难作为出发点。而与卡斯泰尔诺沃同时代的年轻人科拉多·基尼（Corrado Gini）则带领罗马中央统计研究所（Istituto Centrale Statistica in Rome）进行了在精算方面的深入研究。罗马中央统计研究所是一家由保险公司设立的私人研究机构。基尼对所有应用课题的极大兴趣促使他在 20 世纪 30 年代期间与活跃在数理统计领域大部分年轻的意大利数学家保持着密切的联系。

格利文科－坎泰利引理

在这些意大利数学家中有一位叫弗朗切斯科·保罗·坎泰利（Francesco Paolo Cantelli，1875－1966），他差不多先于柯尔莫哥洛夫就建立了概率论的基础。坎泰利对基础理论研究（如研究概率的意义是什么？）不感兴趣，没有像柯尔莫哥洛夫那样更深入地研究概率论，

---

[40] 意大利法西斯主义在最初是非常重视家庭的，因此，只有已婚男子才可以在政府任职，包括在大学里任教。1939 年，非常有才华的布鲁诺·德费奈蒂（Bruno de Finetti）参加了里雅斯特大学（the University of Trieste）一个数学系全职教授职位在全国范围内的竞争，虽然胜出，但是，因为他当时还是一个单身汉，所以没有资格得到这个职位。

他只是满足于用概率运算的各种方法去推导出一些基本的数学定理，而这些概率运算的方法都是自 18 世纪数学家亚伯拉罕・棣莫弗将微积分引入概率计算后就存在的。1916 年，坎泰利发现了我们所称的数理统计的基本原理。尽管它非常重要，却起了一个不起眼的名字“格利文科－坎泰利引理”（the Glivenko-Cantelli Lemma）[41]。坎泰利是第一个证明了这个定理的人，并且，他非常理解它的重要性。至于柯尔莫哥洛夫的学生——约瑟夫・格利文科（Joseph Glivenko）对此定理也做出了贡献，他采用一种新的数学符号，即斯蒂尔切斯积分（Stieltjes integral）概括了这一结果，他的论文在 1933 年发表于一本意大利的数学期刊。格利文科所采用的数学符号是现代教科书中使用最多的一个符号。

格利文科－坎泰利引理是那种直观上显而易见的，但是，只有当别人发现后，你才会意识到，否则看不出来。如果有一些数，我们对它们的概率分布一无所知，那么数据本身可以用来构造一个非参数分布，这是一个不那么好看的数学函数，其间有许多断点，怎么看都不优美，尽管它的结构不雅观，坎泰利还是可以通过增大观测值的数量，来使不那么美的经验分布函数（empirical distribution function）越来越接近真实的分布函数。

格利文科－坎泰利引理的重要性立刻得到了承认，在这之后的 20 年里，这个引理被用来还原并证明了许多重要的定理，它是一种经常用于证明中的数学研究工具之一。为了用这个引理，数学家在 20 世纪初，不得不想出一些计算方法的简便算法，如果没有小窍门，在大量的数据样本中用经验分布函数来进行参数估计，就需要有一部在一秒钟内可以进行数百万次计算的超强计算机。在 20 世纪 50 年代、60 年代乃至 70 年代都还没有这样的机器，到了 80 年代，才有这样的计算机用于这样的计算。格利文科－坎泰利引理成为新统计方法的基础，而这种新统计方法只能生存在高速计算机的世界里。

埃弗龙的“解靴带”法

在 1982 年，斯坦福大学的布拉德利・埃弗龙（Bradley Efron）发明了所谓“解靴带”（Bootstrap）（我们称为“自助法”）的方法，它基于格利文科－坎泰利引理的两种简单应用。这两种应用方法的原理很简单，但是它们要求用电脑进行大量的计算、再计算，……如果对一组数量适中的数据进行典型的“解靴带”分析，即使是利用最好的计算机也需要花好几分钟的时间。

埃弗龙把这种方法称为“解靴带”，是因为整个计算过程是一个数据自身模拟提升的过程，就像是解靴带一样，一个接一个地被解开。计算机不会介意重复单调的工作，它一遍又一遍地做着同样的工作，从不抱怨。由于使用了现代的晶体管芯片，计算机可以在不到万分之一秒内完成这些工作。在埃弗龙的“解靴带”背后还有一些复杂的数学理论，他最初的论文中证明了，如果对真实的数据分布做出了恰当的假设，这个方法与标准方法是等同的。这个方法的应用非常广泛，从 1982 年开始，几乎在每个数理统计期刊上都刊载一篇或更多的与“解靴带”相关的文章。

重复抽样和其它运算密集方法

还有其它一些与“解靴带”类似的方法，总称为重复抽样（resampling）。事实上埃弗龙已经阐述了费歇尔的许多标准统计方法都可以看作是重复抽样，而且，重复抽样方法属于范围更广的统计方法的一种，我们称之为“运算密集”（computer-intensive）。运算密集法充分利用现代计算机，对相同的数据不断地重复进行大量的运算。

20 世纪 60 年代，美国国家标准局（the National Bureau of Standards）的琼・罗森布拉

---

[41] 在 18 世纪，欧几里德的《几何原本》（Elements）的形式数学（formal mathematics）被编入几何学教科书，而且将逻辑推理的模式也编进去了。在编纂过程中，“定理”（theorem）一词被用来描述手头特定问题的结论。而为了证明某些定理，我们必须先去证明在最终定理中要用到的一些过程结论，当然，这些过程结论也可用来证明其它定理。这样的过程结论称为“引理”（lemma）。

特（Joan Rosenblatt）和德州农工大学（Texas A&M University）的伊曼纽尔•帕仁（Emmanuel Parzen）发展了这种运算密集的程序，他们的方法被称为“核密度估计”（kernel density estimation），而且，由此产生了“核密度回归估计”（kernel density-based regression estimation）。这两种方法涉及到两个任意参数，一个是“核”（kernel），另一个是“带宽”（bandwidth）。这些方法出现不久，1967 年（远在计算机可以解决这些问题之前）哥伦比亚大学的约翰•范里津（John van Ryzin）利用格利文科－坎泰利引理确定了参数的最优配置。

当数理统计学家们还在研究理论，并在他们自己的期刊发表文章时，罗森布拉特和帕仁的核密度回归已经被工程界独立地发现了，在计算机工程师中，它被称为“模糊近似值”（fuzzy approximation）。它用了范里津所称的“非最优核”（nonoptimal kernel），并且，只是非常随意地选用了一个“带宽”。工程实践不是为了寻找理论上最佳的可能方法，而是在于追求可行性。当理论家们还在为抽象的最优标准而大费周折时，工程师们已经走出去，到了真实的世界，用模糊近似值的概念建立了以计算机为基础的模糊系统。模糊工程系统应用于傻瓜相机，可以自动对焦和调整光圈。这一系统还应用于新建筑物中，根据不同房间的不同需要调整和保持舒适的恒定室温。

巴特•科什科（Bart Kosko）是工程界一个私人咨询师，是模糊系统推广者中最成功的一位。当我读他书中列出的参考书目时，可以找到关于 19 世纪一些主流数学家，像戈特弗里德•威廉•冯•莱布尼茨（Gottfried Wilhelm von Leibniz）等的参考资料，还有对随机过程理论及其在工程领域的应用方面做出贡献的数理统计学诺伯特•维纳（Norbert Wiener）的一些资料。但我找不到罗森布拉特、帕仁、范里津或核回归理论（the theory of kernel-based regression）任何后来贡献者的资料。这表明，尽管模糊系统和核密度回归的计算机运算法则基本一致，但它们各自完全独立地得到了发展。

统计模型的胜利

运算密集法在标准工程实践中的扩展，是 20 世纪末统计革命已经渗透到科学界各个角落的一个实例。数理统计学家们已经不再是统计方法发展唯一的、甚至已经算不上是最重要的参与者了。在过去的 70 年中，科学家和工程师们并不知道那些刊载于他们期刊中最重要的理论经常一次次地被重新发现[42]。

有时，应用者应用基础定理时没有进行重新论证，仅仅凭直觉上以为是对的就假定它是正确的。还有的情况是，使用者使用了已经被证明是错误的定理，仅仅是因为这些定理直观上看起来是正确的。存在这种问题的原因，是因为在现代科学教育中概率分布的概念已经根深蒂固，以至于统计学家和工程师们思考问题的方式也是基于概率分布的角度。一百多年前，K•皮尔逊认为，所有的观测都来自于概率分布，而科学的目的就在于估计这些分布的参数。在这之前，科学界相信宇宙遵守着某些规律，如牛顿运动定律，而观测到的任何差异都是因为误差的存在。逐渐地，皮尔逊的观点占据了优势，其结果，每个在 20 世纪接受科学方法训练的人都理所当然地接受了皮尔逊的观点。这种观点深深地植根于现代数据分析的科学方法之中，几乎没有人去考虑其所以然。很多科学家和工程师使用这些方法，但从不考虑 K•皮尔逊观点的哲学含义。

然而，当科学研究的真正“主体”是概率分布这一观念被广为接受时，哲学家和数学家发现了许多严重的基本问题，我已经在以上的章节中概略地列举了一些，在下一章节将详细论述。

---

[42] 我在我的博士论文中，使用了一种众所周知的、至少统计学家们都称之为“复合泊松分布”（compound-Poisson distribution）的分布。当我写论文时我必须去查相关的资料，结果我发现了在经济学、运筹学、电子工程以及社会学中都有同样的分布。有些地方它被称为“结巴泊松”（stuttering Poisson）或泊松二项分布（Poisson-binomial）。在一篇论文中，它还被称为“第五街公共汽车分布”（Fifth Avenue bus distribution）。

第 29 章 “泥菩萨”

1962 年，芝加哥大学的托马斯・库恩（Thomas Kuhn）出版了《科学革命的结构》（The Structure of Scientific Revolutions）一书。这本书深刻地影响了哲学家们和实践者们如何去看待科学。库恩指出，现实是复杂的，是绝对不可能由一个有组织的科学模型来完全描述出来的。他认为科学就是试图模拟建立一个描述现实的模型，符合可用的数据，并且可以用来预测新实验的结果。因为没有任何一个模型是完全真实的，所以，数据越来越多，要求不断地配合新的发现去修正模型以修正对现实的认知。这样，模型因为带有特例的直觉上难以置信的延伸，变得越来越来复杂，最终，这个模型不再适用了。这时，有创新精神的人将会考虑建立一个全新的模型，一场新的革命在科学领域即将展开。

统计革命就是模型变换的例子。用 19 世纪决定论的科学观，牛顿物理学已经成功地描述了行星、月球、小行星和彗星等天体的运动，运动都是遵守几个明确的运动和引力定律；在寻找化学规律方面也取得了一些成功；并且达尔文的自然选择学说为理解进化提供了有利的依据；甚至有些人试图将这种寻找科学规律的模型研究引入社会学、政治科学以及心理学等领域。那时，人们相信寻找规律的难点在于测量不准确。

19 世纪初，一些数字家如皮埃尔・西蒙・拉普拉斯认为，天文测量存在微小误差，可能是因为大气状况和测量的人为因素。他提出，这些误差也应该存在一个概率分布，从而开启了统计革命的大门。按照库恩的观点，这就是在获得新的数据后对机械式宇宙观进行的修正。19 世纪，比利时学者兰伯特・阿道夫・雅克・凯特莱（Lambert Adolphe Jacques Quételet）最早开创了统计革命，他认为人类行为的规律也具有概率论的性质。他没有用皮尔逊的多参数方法，并且也不知道最佳估计方法（optimum estimation），他的模型是极其朴素的。

最终，人们发现，更加精确的测量反倒使模型预测值和实际观测值之间的差异变得更大，关于科学的决定论观点彻底崩溃，测量的越加精确，不但没有按照拉普拉斯的想法去消除误差，反而降低了人们观测行星真实运动的能力，而且表现出的差异越来越大。基于这一点，科学界已经做好了接受皮尔逊及其参数分布的准备。

本书前面的章节已经介绍了皮尔逊的统计革命是怎么逐渐改变整个现代科学的，尽管分子生物学遵循这种决定论（基因会决定细胞产生特殊的蛋白质），但是，在该科学中产生的实际数据充满了随机性，而且基因事实上就是这些随机数据分布的参数。现代药物对人体功能的影响是绝对的，1 毫克或 2 毫克药物就可能对血压或精神有很大的影响，这一点是确定无疑的。但是证明了这一影响力的药理研究过程，却是按照概率分布来设计和分析的，影响力就是这些分布的参数。

同样，经济计量学的统计方法被用来模拟一个国家或者一个企业的经济活动。我们确信的电子的质子这些次原子粒子在量子力学中都是作为概率分布描述的。社会学家用总体的加权算术平均数来描述个体的交互作用，但这只能按照概率分布的方式进行。在许多类似的科学领域里，统计模型的应用在它们的方法论中非常广泛。当谈及分布的参数时，好像它们是真的并且是可测量的一样。多变且不确定的数据集合，就是这些科学的起点，计算结果则是隐藏在大量计算中，以参数形式来表示，这些参数是永远不能通过直接观测得到的。

统计学家失去控制权

现代科学中的统计革命如此彻底，以致于统计学家已经失去了对过程的控制。在数理统计文献的基础上，分子遗传学家已经独立发展了自己的概率计算方法。计算机对大量数据

的处理能力，和人们对整理并搞清楚这些巨大信息库含义的需求，促使信息科学这一新学科的诞生。在信息科学新期刊的文章中已经很少提到数理统计学家的工作，而且，在《生物统计》或《数理统计年报》中刊登过的许多分析方法，都正在被重新发现。统计模型在公共政策问题研究中的应用，已经演变成了一个被称为“风险分析”（risk analysis）的新学科，并且风险分析的新期刊也忽视数理统计学家的工作。

现在几乎所有新学科的期刊，要求在结论中有一个结果表，列出对统计结论产生影响的不确定因素的测量值。统计分析的标准方法已经成为大学中这些学科的研究生课程，通常，课程的讲授还不必同一个学校的统计系参与。

自 K·皮尔逊发现偏斜分布的一百多年里，统计革命不仅扩展到大多数的科学领域中，而且其许多思想已经传播到了一般的文化当中。当电视新闻主持人宣布，某项医学研究已经表明被动吸烟的人的死亡风险比不吸烟的人高一倍时，几乎每个听众都认为他或她明白主持人的意思；当一个公众民意调查说 65%的公众对总统表示满意，上下误差 3%时，我们大多数人都认为我们都明白这个 65%和 3%的含义；当我们听到气象播报员预测明天下雨的概率为 95%时，大多数人出门都会带上一把雨伞。

除了这些我们自以为理解的可能性和比例问题外，统计革命对流行思潮和文化，有更深刻的影响力。即使实际测量的数据不够精确地与这些结论吻合，我们还是接受基于估计参数的科学研究结果。我们愿意根据众多数据算出的数来制定公共政策和安排我们的个人计划。我们认为搜集人口出生和死亡的数据，不仅是一个正当的程序，更有必要的工作，我们不必担心数人数会惹怒了上帝。从语言描述方面，我们用“相关”（correlation）或“相关的”（correlated）这两个词，好像它们意味着什么，也好像我们知道其含义。

写这本书的初衷是为了向那些没有数学专业背景的人士解释这场统计革命，我已经尽力描述了在这场革命背后的基本思想，它将如何应用于其他科学领域？它将如何最终主导几乎所有科学领域？我也尽力用语言和实例解释了一些数学模型，使大家不用再去研究抽象的数学符号就能够理解。

统计革命走到尽头了吗？

深邃未及的这个世界是一个集情感、事件与骚动的复杂混合体。我同意库恩的观点，我不相信人类的头脑能够构造一个理想的结构去解释、甚至不能挖地描述这个世界的真实情况。任何这种努力都存在根本的缺陷，最终，这些缺陷会变得非常明显，以至于科学模型必须不断地被修正，最终将走到它的终点，取而代之的是其它的什么东西。

随着统计方法应用的扩展，越来越多地应用到了人类生活的很多领域，哲学问题就显现出来。因此，我认为以讨论哲学问题作为本书的结尾是个好主意。接下来的将是在哲学领域中的一次冒险经历。读者可能想知道哲学究竟对科学信现实生活起到了什么作用。我的答案是，哲学并不是一些被称为哲学家的怪人们所做的神秘学术练习，哲学关注的是我们日常文化思想和活动的基本假设（underlying assumption）。我们的世界观来自于我们的文化，是受许多微妙的假设影响的，甚至很少有人会意识到它们。学习哲学会让我们揭开这些假设，并去检查它们的有效性。

我曾经在康涅狄格大学的数学系教过一门课程，这门课程有一个正式的名称，但是系里的人却更愿称之为“给诗人开的数学”。这门课只开一个学期，是为艺术专业的学生设计的，目的是向他们介绍基本的数学观念。在学期的开始，我向学生们介绍了 16 世纪意大利数学家吉罗拉莫·卡尔达诺（Girolamo Cardano）的一本书《高等艺术》（Ars Magna），在这本书中，第一次描述了代数的方法。与他的大部头著作相呼应，卡尔达诺在该书的介绍中写道：代数不是新东西。他暗示他不是无知的傻子，他认为自人类产生以来，人类对知识的掌

握一直在减少，亚里士多德所拥有的知识远远要多于卡尔达诺那个时代的任何一个人。他断言不可能有新的知识。然而，由于他的无知，他没能在亚里士多德的著作中找到关于代数思想的参考书目，所以他就把代数——这个看起来像是新东西的概念介绍给读者，他确信一些更加有知识的读者会从古人的著作中找到出处，这看起来是新东西的观念一定会被找出来的。

坐在我教室里的这些学生，生活在一个不同的文化环境中，他们不但相信后人会发现新事物，而且事实上，还鼓励创新。他们被卡尔达诺震惊了。写这些是多么愚蠢的呀！我告诉他们，在 16 世纪的时候，因为当时的一些基本哲学假设，欧洲人的世界观具有局限性，他们的世界观中，一个重要的部分就是人类的堕落以及随之而产生的道德、知识、工业等所有事物的持续退化，这些在当时是如此的真实，以至于很少有人去探寻究竟。

我问学生们，他们的世界观的基本假设中，哪些可能在 500 年后看起来是很荒谬的？他们一个都想不出来。

因为统计革命的表面观念已经传播到现代文化中，越来越多的人相信所谓的真实性，而不考虑它的基本假设，所以，让我们用统计的宇宙观来考虑下面三个哲学问题：

1. 可以用统计模型来做决策吗？
2. 当概率应用于现实生活中时其含义是什么？
3. 人们真的懂得什么是概率吗？

可以用统计模型来做决策吗？

牛津大学的 L •乔纳森 •科恩（L. Jonathan Cohen）是被他称之为“帕斯卡式”（“Pascalian”）观点的尖锐批评家，所谓“帕斯卡式”观点就是认为可以用统计分布去描述现实。1989 年他写了《归纳和概率的哲学导论》（An Introduction to the Philosophy of Induction and Probability）一书，书中他提出了一个关于彩票的悖论，他认为那是康涅狄格州卫斯理大学（Wesleyan University in Middletown Connecticut）的西摩 • 屈贝里（Seymour Kyberg）教授发明的。

假定我们接受假设或者显著性检验的观点，我们赞同如果现实中该假设的相应概率非常小，就可以拒绝这个假设。为了更进一步说明，假设 0.0001 就是一个非常小的概率，让我们组织一次公正的 10000 张彩票的抽彩活动。按这个假设，1 号彩票中奖的概率，我们也可以拒绝这种假设，依次类推，我们可以拒绝类似的任何针对某号彩票的假设。按照这一逻辑规则，如果 A 不为真，B 和 C 都不为真，那么 A、B、C 的集合也不为真。也就是说，按照这一逻辑规则，如果每一张彩票都中不了奖，那么就没有彩票可中奖（而事实却是总会有中奖的彩票）。

在科恩较早写的《可能与可证》（the Probable and the Probable）一书中，基于普遍的法律实践，他提出了这种悖论的一个变形。在习惯法（common law）中，一个涉及民事诉讼的原告提供了“有利”证据，其陈述看起来是真的，那么他就会胜诉，法庭接受原先诉求的概率高于 50%。科恩还提出了一个关于“无票入场者”（gate crashers）的悖论：假设在一个有 1000 个席位的音乐厅里举办一场摇滚音乐会，主办单位只售出 499 张票，但是当音乐会开始的时候，1000 个席位都坐满了，根据英国的习惯法，主办单位有权在音乐会上向每个现场的人收票钱，因为他们每个人无票入场的概率都是 50.1%，这样，虽然音乐厅只有 1000 个席位，但是主办单位却将会有 1499 张门票的收入。

这两个悖论都说明了，以概率为依据所得到的决策是不合逻辑的，逻辑和概率是矛盾的。费歇尔在设计良好的实验基础上，利用显著性检验来证明科学研究中的归纳推理是可取的，但是科恩的悖论则表明，这样的归纳推理是不合逻辑的。杰里 • 科恩菲尔德根据积累的大量证据来判断吸烟会导致肺癌这个说法，但连续的研究表明，除非你假设吸烟是致癌的原

因，否则这个结论是极不可能的。相信吸烟致癌是不合逻辑的吗？

以逻辑推理和统计为基础所得出决策上的不一致，是不能靠在科恩提出的悖论中找到错误的假设来解决的。这种不一致的深层次原因存在于逻辑的含义中（科恩认为概率模型可以由一种我们称为“模型逻辑”（model logic）的复杂数学逻辑结构来代替，但是我认为这个方法会产生更多的问题，比它所解决的问题还要多）。在逻辑上，一个命题是对还是错，我们是完全不同的。但是概率引入的观念却是说一些命题“可能”或者“多数”是对的。就是结果的这一点点不确定性，就使我们在分析原因和结果时，难以应用事物实质蕴涵的冷酷的精确性。在临床实验中，处理这类问题的方法，是把每个临床研究看作是对某个治疗方案的效果提供资料。这些资料的价值取决于这个研究的统计分析，但则无也取决于研究的质量。研究质量这一额外的测量决定了哪些研究对结论起决定作用。但是，质量的概念含糊不清而且难以计算，悖论依然存在，而且吞噬着统计方法的核心。这种不一致的毛病是否需要在 21 世纪发起一场新的革命？

当概率应用于现实生活中时，其含义是什么？

柯尔莫哥洛夫建立了概率的数学定义：概率是一个抽象空间里对一事件集合的一种测量。所有概率的数学特征都可由这个定义导出。当我们希望在现实中使用概率时，我们需要确定眼前特定问题事件的抽象空间。当气象播音员说明天降雨的概率为 95%时，什么是所测量的抽象事件的集合？是指明天要外出的所有的人吗？其中有 95%的人会淋雨？还是指可能逗留在外面的时间？其中有 95%的时间我会淋雨？或是说在一个 1 平方英寸大的地方，有 95%的面积会下雨？当然这些解释都不对，那么到底是什么意思呢？

柯尔莫哥洛夫之前的 K·皮尔逊认为概率分布是可以通过收集到的数据观察得出的，我们已经看到了使用这个方法存在的问题。

威廉·S·戈塞特试图为一个设计好的试验描述其事件空间。他说事件空间就是试验得出所有可能结果的集合。这听起来可能是对的，但是在实践中却是无用的。在实验中，我们必须相当精确地描述出结果的概率分布，才能计算出统计分析中需要用到的概率值。“所有可能实验结果的集合”的概念非常含糊，我们怎样才能得到一个精确的概率分布呢？

起初费歇尔同意戈塞特的想法，继而他发展了一个更好的定义。在他的实验设计中，治疗方案是随机分配给各个实验单位的。如果我们想在肥老鼠身上做实验，比较两个治疗动脉硬化的方案，我们就随机地在一些老鼠身上使用 A 方法，而在其余的老鼠身上使用 B 方法。实验开始进行，我们开始观察结果。假设两种治疗方案具有同样的效果，因为动物是随机使用治疗方法的，所以另外一些分配治疗的效果应该是同样的。随机治疗方法的标签是不相关的，只要治疗效果是一样的，我们就可以在动物间随意调换。因此，对于费歇尔，事件的空间是所有可能随机分配的治疗方案的集合。这是一个事件的有限集合，所有的事件都是等概率发生的。在所有治疗方法的效果是相等的零假设（null hypothesis）条件下，实验结果的概率分布是可以计算出来的，这就是我们所说的排列检验（permutation test）或随机检验。当费歇尔提出这一检验方法时，还不能计算出所有可能的随机实验分配方式，费歇尔证明了，他的方差分析公式可以求得一个非常理想的排列检验的近似值。

那时还没有高速计算功能的计算机，而现在进行排列检验是可能的，因为电脑可以不知疲倦地进行计算，这样费歇尔的方差分析公式就不再需要了，而且很多数理统计学家经过多年求证得出的非常聪明的定理也不再需要了。只要数据结果是来自于一个随机控制的实验，就可以在计算机上用排列检验来进行所有的显著性检验。

如果对观测数据用一个显著性检验，那就不可能了。这是费歇尔反对吸烟与健康问题研究的主要原因。一些论文的作者使用统计检验方法证明他们的例子。费歇尔认为，除非他们研究的是随机化的实验，否则统计显著性检验就是不合适的。在美国法院中的歧视性案件

就常常是根据统计的显著性检验来裁决的。美国最高法院（The U. S. Supreme Court）规定，统计显著性检验是一种可以在裁决中使用的方法，可以用来判定是否因为性别或种族歧视的原因而造成了影响。费歇尔如果知道，他一定会强烈反对。在 20 世纪 80 年代后期，美国国家科学院（The U. S. National Academy of Science）赞助了一项研究，研究在法院中使用统计方法作为裁决依据是否合理。这项研究的主持者是卡内基梅隆大学（Carnegie Mellon University）的斯蒂芬·菲恩伯格（Stephen Fienberg）和明尼苏达大学（the University of Minnesota）的塞缪尔·克里斯洛夫（Samuel Krislov）。这个研究小组在 1988 年发表了他们的研究报告。研究报告中的许多论文批判了将显著检验用于歧视性案件的作法，所持的论点类似于费歇尔在反对吸烟导致癌症的证据时所使用的理由。如果最高法院想在诉讼中使用显著性检验，它必须确定产生概率的事件空间。

如何找出柯尔莫哥洛夫事件空间？第二种方法来自于样本调查理论。当我们希望通过一个随机样本去判断整个群体的某些事时，我们要精确地确定要研究的人群总体，确立一个选取样本的方法，并且根据该方法进行随机抽样。在实验的结论中存在不确定性，我们可以使用统计方法来量化这一不确定因素。不确定性产生的原因，是因为我们处理的是样本而不是所有人群。我们研究的宇宙现象的真实数值是固定不变的，例如，支持总统施政政策的美国选民的百分数是确定的，只是他们不知道。能够使用统计方法的事件空间，是所有可能的随机样本的集合，同样，这是一个有限集合，它的概率分布是可以计算出来的。概率在现实生活中的含义清楚地建立在抽样调查之上。

当统计方法应用于天文学、社会学、流行病学、法律或者天气预报等观测研究中时，事件空间就不好确定。在这些领域之中的很多争论，通常都是因为不同的数学模型会产生不同的结论。如果我们不能确定可进行概率计算的事件空间，那么就不能说某种模型比另外一种更适用。就像在很多法律案件中所显示的那样，两个统计专家分析同一组数据却得不到统一的结论。当统计方法越来越多地被政府和社会团体应用到观察研究和解决社会问题时，这个基本问题的存在，即不可能算出确切概率的事实，将使人们对这些统计方法的有效性产生怀疑。

人们真的懂得什么是概率吗？

概率在现实生活中还有一个含义是“个人概率”。美国的 L·J·萨维奇和意大利的布鲁诺·德费奈蒂是倡导这种观点的先驱。其先驱地位的确定是因为萨维奇 1954 年出版的《统计学基础》（*The Foundations of Statistics*）一书。在这种观点下，概率是一个广泛的概念，人们很自然地使用概率来支配生活。在进行冒险前，人们总会本能地根据可能产生结果的概率根据可能产生结果的概率进行决策，如果预想危险的概率很高，人们就会采取回避的态度。对萨维奇和德费奈蒂来说，概率是一个普通的概念。人们不必去联系柯尔莫哥洛夫的数学概率，我们所要做的就是建立一些一般性的规则，将个人概率与生活联系起来，因此，我们只要假设人们在判断事件的概率时所遵照的规则是一致的就可以了。萨维奇在这一假设下提出了一些关于内部一致性的规则。

按照萨维奇和德费奈蒂的方法，个人概率对每个人来讲是独特的。对同样的数据进行同样的观察，有的人会判断降水概率是 95%，有的人则会判断是 72%，这样的事情是极有可能发生的。利用贝叶斯定理，萨维奇和德费奈蒂向人们展示了具有相同个人概率的两个人如果分析的是同一序列数据，最终他们会得到相同的概率估计。这是一个令人满意的结论：人看起来都是不同的，但却都是理性的。如果提供了足够的数据，理性的人们会最终求得共识，哪怕最初他们是存在意见分歧的。

约翰·梅纳德·凯恩斯在 1921 年发表的题为《关于概率的讨论》（A Treatise on Probability）的博士论文中，对个人概率提出了不同的看法。凯恩斯认为，概率是在某一文化教育背景下

的人们，对其既定情况的不确定性的测量，概率的判断不仅是个人内心的直觉，还与个人的文化背景有关系。如果我们想在 72%和 68%之中作出哪一个更准确的选择，用凯恩斯的方法就会很困难，因为人们的总体文化水平很难达到精确的同一程度。凯恩斯指出，如果只是为了做决定，我们很少或根本不必去知道这些事件确切的概率数值，只要将事件进行排序就足够了。根据凯恩斯的理论，我们只要知道哪一事件更可能发生就可以了。明天下雨比下冰雹的可能性要大，或者说明天下雨的可能性是下冰雹可能性的两倍。凯恩斯指出，概率可以是部分排序（partial ordering）。不必要把每件事与其它事情进行比较。我们可以忽视某些概率关系，如根本不必要把扬基队得总冠军的概率与明天下雨的概率联系起来。

照这样，关于概率含义的两个结论取决于人类对不确定性量化的愿望，或者至少是大致的量化的要求。在凯恩斯的《关于概率的讨论》中，他为他的个人概率的部分序列设计出了一个正式的数学结构。他的做法比柯尔莫哥洛夫为数学概率建立基础理论还要早。他所做的工作没有借鉴柯尔莫哥洛夫的理论。凯恩斯声称，他的概率的定义有别于 1921 年提出的概率数学的一系列数学计算公式。为了使凯恩斯的概率定义得到应用，使用者还必须符合萨维奇的一致性原则。

凯恩斯的定义提供了关于概率的一种观点，它是用统计方法进行决策的基础。这种观点认为概率不再以事件空间为基础，而是产生于所涉及人员的个人感觉的数值。接着希伯来大学（Hebrew University）的两个心理学家——丹尼尔·卡内曼（Daniel Kahneman）和阿莫斯·特韦尔斯基（Amos Tversky）开始了他们关于个人概率的心理学研究。

在 20 世纪 70 年代和 80 年代间，卡内曼和特韦尔斯基研究了个体理解概率的方式。他们的研究成果编入了由 P·斯洛维奇（P. Slovic）编辑的《不确定情况下的判断——启发与偏见》（Judgment under Uncertainty: Heuristics and Biases）一书中。他们为大学生、大学教员和一般的市民提出了许多概率场景，他们发现没有人符合萨维奇的一致性原则，相反，大多数人对不同概率数值的含义甚至没有一个一致的观点。他们所发现最好的一点就是人们对 50：50 和“几乎肯定”的含义有着一致的认识。通过卡内曼和特韦尔斯基的研究，我们可以得出结论：天气预报员尽力想区分降雨概率 90%和 75%间的不同，但实际上他们根本不可能说清楚，而那些预报的收听者也不可能真的说清楚这两者间的区别。

1974 年，特韦尔斯基在皇家统计学会的一次会议上宣布了他的研究结果。在随后的讨论中，斯坦福大学的帕特里克·苏佩斯（Patrick Suppes）提出了一个简单的概率模型，符合柯尔莫哥洛夫的公理，并且也模拟卡内曼和特韦尔斯基的发现。这意味着用这个模型的人在他们的个人概率方面应该是一致的，在苏佩斯的模型中只有五个概率值：

必然为真

为真的可能性大

为真的概率为一半

为真的可能性小

必然为假

这导出了一个很无趣的数学理论。大概只有六个理论可由此模型导出，并且它们的论证几乎是不言而喻的。如果卡内曼和特韦尔斯基是对的，那么惟一有用的个人概率将对奇妙的抽象数学理论十分不利，并且由此产生的统计模型极基有限。事实上，如果苏佩斯的模型是惟一适合个人概率的模型，许多标准统计分析方法就毫无用处了，因为它们算出的差异水平低于人类感觉的水平。

概率真的必要吗？

统计革命背后的基本观点是：科学真实的主体是数字的分布，这个分布可以通过参数来描述。将概念溶入概率理论并处理概率分布，这是数学的方便之处。将数字的分布看作是

概率数学理论的元素，这样就可以建立参数估计量的最优化标准，然后，去解决用数据描述分布时遇到的数学问题。因为概率看起来与分布的概念的关系是与生俱来的，许多人做了很多工作，试图让人们理解概率的含义，努力将概率的含义与现实生活联系起来，并且使用条件概率这一工具去解释学实验和观测的结果。

分布的思想可以存在于概率理论之外。事实上，许多“非正常分布”（improper distributions）（因为这些分布不符合概率分布的所有要求）已经应用于量子力学和一些贝叶斯方法中。排队论（queuing theory）（指两次排队间的平均间隔时间等于在队伍中等候的平均时间）的发展，推导出一个非正常的分布——描述一个人加入队伍必须要等候的时间。这正是一个将概率论的数学理论应用于实际生活，同时却将我们带离概率分布集合的一个例子。

21 世纪将会发生什么事？

柯尔莫哥洛夫表现出来的最后的聪明才智，是他用一组有限符号序列的特性来描述概率。在这个描述中，信息理论不是概率计算的结果，而是概率本身的起源。也许在将来，某个人会继续他的工作，并且发展一个新的分布理论，而在新的分布理论中数字计算机的特性会被带入哲学理论的范畴。

谁知道呢？也许在什么地方有另外一个费歇尔，正工作于科学的最前沿，并在不久的将来，会以其前所未有的见识和观念打破目前的书面？也许在中国的内地，另一个吕西安•勒卡姆已经在一个没有文化的农家出生了；或者在北美，另一个乔治•博克斯只上了初中就休学了，现在正在做机修工，正在努力自学；也许另一个格特鲁德•考克斯将要放弃当传教士的愿望，被科学和数学的谜团深深吸引；或者另一位威廉•S•戈塞特正在努力寻找方法去解决啤酒发酵问题；或者另一个奈曼或皮特曼正在印度某个偏远的地方学院里教书，并且思考着深奥的问题。谁知道下一个伟大的发现将发生在什么地方？

当我们进入 21 世纪的时候，统计革命在科学领域取得了胜利，除了极少数的角落，它已经征服了科学界几乎所有领域的决定论观点。统计观点的应用如此广泛，以至于其基本假设已经成为西方世界通俗文化的一部分，就如同一尊泥菩萨一样立在那里，洋洋得意，而在未来的某个隐蔽的角落，另一场科学革命正在孕育，而那些即将发起这场革命的男男女女，可能正生活在我们中间。

作者后记

在写这本书之前，我已经将那些对统计发展有贡献的女士和先生们分成了两组，一组是我在书中提及到的，一组是我没有提及的。第一组人可能对我在书中只提及他们一小部分的工作而感到不满意，第二组人可能会因为我根本就没有提及他们的工作而表示抗议。。为了表达我对他们的敬意，我有必须解释一下我取舍的原则。

对第一组取舍的原因在于：现代科学的范畴太大了，任何人都不可能知道它所有的支派。因此，在有些研究领域，统计方法的应用可能非常广泛，但是我却不知道。在 20 世纪 70 年代，我曾查找过关于计算机在医学诊断中应用的资料。在查找过程中，我发现有三个互相独立的支派，在任何一个支派内人们互相引述论文，并且都发表在同一份期刊内，但是，不同派别的科学家却很少了解其他派别的人在做什么。这还只是在医学界这样一个小小的相关领域中的情形，在更广阔的科学界，可能有很多人群在应用统计方法，并且可能有一些成果在我从来没听过的期刊中发表。我对统计革命结果的认识，来自于对一些数理统计主流期刊的阅读。不阅读这些主流期刊或者不在这些期刊中发表文章的统计学家，就像发展模糊集合论（fuzzy set theory）的工程师，他们可能做了很多值得记载的工作，但是因为他们不在我知道的科学或数学期刊上发表文章，那么他们的工作就不会被包括进来。

有些东西我是知道的，但还是被省略了。我不想写一本关于统计方法论发展的全面的历史书，因为这本书的读者定位是一些不懂或者略懂数学的人，所以我不得不选择一些能用文字而不是用数学符号来解释的例子，这就更进一步限定了我的选择。另外，我还想让这本书读起来比较流畅，如果我用了数学符号，我可能就可以说明了众多主题间的关系了。但是没有数学符号，这本书很容易退化为一种观念的介绍，这些观念间没有什么关系。这本书需要一条主线将各个主题组织起来，我所选择的贯穿 20 世纪统计学复杂理论的主线是与别人不一样的，一旦这条主线确定了，我就不得不忽视了统计学的很多方面，而实际上，我对它们同样非常感兴趣。

在我的书中，很多人我都没有提及到，这并不代表他们的工作不重要，更不代表我认为他们的工作不重要。仅仅是因为本书的结构限制，我没有办法将他们的研究写进来，只好放弃。

我希望读者读了本书后能有所启发，去进一步了解统计革命的内涵。我希望有人在读后甚至能钻研这个题目，加入统计研究的行列。在参考书目中，我选择了一些供没有数学学习背景的人阅读的图书和文章。在这些书中，其他许多统计学家尝试向我们解释了统计所学带给他们的乐趣，那些想进一步了解统计革命的读者将会喜欢其中的一些书。

我要感谢 W. H. Freeman 出版的公司相关人员在本书出版过程中所做的工作。感谢 Don Gecewicz 细致的校对与编辑；感谢 Eleanor Wedge 和 Vivien Weiss 最后文字定稿和进一步的校对；感谢 Patrick Farace 对本书潜在价值的肯定；感谢 Victoria Tomaselli、Bill Page、Karen Barr、Meg Kuhta 和 Julia Derosa 对本书的美术制作工作。

大事年表

年份	事件	人物
1857	卡尔·皮尔逊出生	K·皮尔逊（Karl Pearson）
1865	圭多·卡斯泰尔诺沃出生	G·卡斯泰尔诺沃（Guido Castelnuovo）
1866	格雷戈尔·门德尔从事植物杂交实验	G·门德尔（Gregor Mendel）
1875	弗朗切斯科·保罗·坎泰利出生	F·P·坎泰利（Francesco Paolo Cantelli）
1876	威廉·西利·戈塞特出生	W·S·戈塞特（“学生”）（William Sealy Gosset）
1886	保罗·利维出生	P·利维（Paul Lévy）
1890	罗纳德·艾尔默·费歇尔出生	R·A·费歇尔（Ponald Aylmer Fisher）
1893	普拉桑塔·钱德拉·马哈拉诺比斯出生	P·C·马哈拉诺比斯（Parasanta Chandra Mahalanobis）
1893	哈拉尔德·克拉美出生	H·克拉美（Harald Cramér）
1894	耶日·奈曼出生	J·奈曼（Jerzy Neyman）
1895	发现偏斜分布	K·皮尔逊
1895	埃贡·S·皮尔逊出生	E·S·皮尔逊（Egon S. Pearson）
1899	切斯特·布利斯出生	C·布利斯（Chester Bliss）
1900	格特鲁德·M·考克斯出生	G·M·考克斯（Gertrude M. Cox）
1900	重新发现格雷戈尔·门德尔的成果	W·贝特森（W. Bateson）
1902	《生物统计》（Biometrika）第 1 期出版	F·高尔顿（F. Galton）、K·皮尔逊、R·韦尔登（R. Weldon）
1903	安德烈·尼古拉耶维奇·柯尔莫哥洛夫出生	A·N·柯尔莫哥洛夫（Andrei Nikolaevich Kolmogorov）
1906	塞缪尔·S·威尔克斯出生	S·S·威尔克斯（Samuel S. Wilks）
1908	《平均数的可能误差》（“The probable Error of the Mean”）“学生”t 检验（student’s t-test）	W·S·戈塞特
1909	弗洛伦斯·南丁格尔·大卫出生	F·N·大卫（Florence Nightingale David）
1911	弗朗西斯·高尔顿爵士去世	F·高尔顿（Francis Galton）
1911	《科学的法则》（The Grammar of Science）	K·皮尔逊
1912	杰尔姆·科恩菲尔德出生	J·科恩菲尔德（Jerome Cornfield）
1912	费歇尔发表第一篇论文	R·A·费歇尔
1915	相关系数（correlation coefficient）的分布	R·A·费歇尔
1915	约翰·图基出生	J·图基（John Tukey）
1916	格利文科—坎泰利引理（Glivenko-Cantelli lemma）首次出现	F·P·坎泰利
1917	L·J·萨维奇出生	L·J·萨维奇（L. J.（“Jimmie”）

		Savage）
1919	《概率运行与应用》（Calcotlo della probabilità…）出版	G •卡斯泰尔诺沃（G. Castelnuovo）
1919	费歇尔在罗森斯特实验站（Rothamsted Experimental Station）	R • A • 费歇尔
1920	关于勒贝格积分（Lebesgue integration）的第一篇论文发表	H • 勒贝格（H. Lebesgue）
1921	《关于概率的讨论》（A Treatise on Probability）	J • M • 凯恩斯（J. M. Keynes）
1921	《作物收成变动研究Ⅰ》（Studies in Crop Variation. Ⅰ）	R • A • 费歇尔
1923	《作物收成变动研究Ⅱ》（Studies in Crop Variation. Ⅱ）	R • A • 费歇尔
1924	《作物收成变动研究Ⅲ》（Studies in Crop Variation. Ⅲ）	R • A • 费歇尔
1924	《消除心智缺陷》（The Elimination of mental Defect）——费歇尔关于优先学的第一篇文章	R • A • 费歇尔
1925	《研究工作者的统计方法》（Statistical Methods for Research Workers）第一版出版	R • A • 费歇尔
1925	统计估计理论（极大似然估计（ML Estimation））	R • A • 费歇尔
1926	关于农业实验设计的第一篇论文	R • A • 费歇尔
1927	《作物收成变动研究Ⅳ》（Studies in Crop Variation. Ⅳ）	R • A • 费歇尔
1928	奈曼－皮尔逊（Neyman－Pearson）关于假设检验（hypothesis testing）的第一篇论文	J • 奈曼、E • S • 皮尔逊
1928	三条极值渐近线	L •H •C •蒂皮特（Tippett）、R •A •费歇尔
1928	《作物收成变动研究Ⅵ》（Studies in Crop Variation. Ⅶ）	R • A • 费歇尔
1930	《数理统计年报》（Annals of Mathematical Statistics）第一期出版	H • 卡弗（H. Carver）
1930	《自然选择的遗传理论》（The Genetical Theory of Natural Selection）	R • A • 费歇尔
1931	印度统计研究所（Indian Statistical Institute）成立	P • C • 马哈拉诺比斯（P. C. Mahalanobis）
1933	概率的公理化	A • N • 柯尔莫哥洛夫
1933	《印度统计年报》（Sankhya）第一期出版	P • C • 马哈拉诺比斯
1933	概率单位分析（probit analysis）成果完成	C • 布利斯（C.Bliss）
1933	塞缪尔 • S • 威尔克斯到达普林斯顿（Princeton）	S •S •威尔克斯（Samuel S. Wilks）
1934	奈曼的置信区间（confidence intervals）	J • 奈曼
1934	中心极限定理（central limit theorem）的证明	P • 利维、J • 林德伯格
1934	切斯特 • 布利斯在列宁格勒植物保护研究所（Leningrad Institute for Plant Protection）	C • 布利斯（Chester Bliss）
1935	鞅理论（martingale theory）的首次发展	P • 利维

1935	《实验设计》（The Design of Experiments）出版	R・A・费歇尔
1936	卡尔・皮尔逊去世	K・皮尔逊
1937	利用随机抽样对美国失业普查进行数字检查	M・汉森（M. Hansen）、F・斯蒂芬（F. Stephan）
1937	威廉・西利・戈塞特去世	W・S・戈塞特（“学生”）
1938	《生物、农业与医疗研究统计表》（Statistical Tables for Biological，Agricultural，and Medical Research）	R・A・费歇尔、F・耶茨（F. Yates）
1940	《统计方法》（Statistical Methods）教科书	G・W・斯内德克（G. W. Snedecor）
1941	亨利・勒贝格去世	H・勒贝格（Henri Lebesgue）
1945	在《统计的数学方法》（Mathematical Methods of Statistics）中对费歇尔的成果进行修订	H・克拉美
1945	威尔科克森关于非参数检验的第一个出版物	F・威尔科克森
1947	在出版物中第一次出现序贯估计理论（sequential estimation theory）	A・沃尔德
1947	曼一惠特尼（Mann一Whitney）对非参数检验的表述	H・G・曼、D・R・惠特尼
1948	皮特曼在非参数统计推断方面的成果	E・J・G・皮特曼（E. J. G. Pitman）
1949	科克伦关于观测研究的成果	W・G・科克伦（W. G. Cochran）
1950	科克伦和考克斯关于实验设计的著作出版	W・G・科克伦、G・M・考克斯
1952	圭多・卡斯泰尔诺沃去世	G・卡斯泰尔诺沃（Guido Castelnuovo）
1957	费歇尔关于吸烟假定危险的辩论	R・A・费歇尔
1958	《极值统计学》（Statistics of Extremes）出版	E・J・冈贝尔（E・J・Gumbel）
1959	博克斯使用“稳健”（“rebust”）这一术语	G・E・P・博克斯（G. E. P. Box）
1959	假设检验的最终表述	E・L・莱曼（E. L. Lehmann）
1960	《组合机遇》（Combinatorial Chance）	F・N・大卫、D・E・巴顿（D. E. Barton）
1962	萨维奇一德费奈蒂（Savage-de Finetti）个人概率理论的表述	L・J・萨维奇、B・德费奈蒂（B. de Finetti）
1962	费歇尔关于遗传学中性别差异的最后论文	R・A・费歇尔
1962	罗纳德・艾尔默・费歇尔去世	R・A・费歇尔
1964	塞缪尔・S・威尔克斯去世	S・S・威尔克斯
1964	《变换分析》（An analysis of transformations）	G・E・P・博克斯、D・R・考克斯（D. R. Cox）
1966	弗朗切斯科・保罗・坎泰利去世	F・P・坎泰利
1967	哈耶克秩检验的表述	J・哈耶克（J. Hájek）
1969	全国性三氟溴氯乙烷研究（包括对数线性模型的成果）	Y・M・M・毕晓普（Y. M. M. Bishop）及其他人
1970	南希・曼关于可靠性理论（reliability theory）和威布尔分布（Weibull distribution）的第一个出版物	N・曼（Nancy Mann）
1970	《赛局、上帝与赌博》（Games，Gods，and Gambling）	F・N・大卫
1971	保罗・利维去世	P・利维
1971	L・J・萨维奇去世	L・J・萨维奇

1972	普林斯顿稳健估计研究（普林斯顿稳健性研究）	D・F・安德鲁（D. F. Andrews）、P・J・比苛尔（P. J. Bickel）、F・R・汉佩尔（F. R. Hampel）、P・J・休伯（P. J. Huber）、W・H・罗杰斯（W. H. Rogers）、J・W・图基（J. W. Tukey）
1972	普拉桑塔・钱德拉・马哈拉诺比斯去世	P・C・马哈拉诺比斯
1975	斯特拉・坎利夫当选皇家统计学会（Royal Statistical Society）会长	S・V・坎利夫（Stella Cunliffe）
1976	“科学与统计学”，显著性检验应用的一个观点	G・E・P・博克斯
1977	考克斯对显著性检验的表述	D・R・考克斯
1977	《探索性数据分析》（Exploratory Data Analysis）出版	J・图基
1978	格特鲁德・M・考克斯去世	G・M・考克斯
1979	切斯特・布利斯去世	C・布利斯
1979	杰尔姆・科恩菲尔德去世	J・科恩菲尔德
1979	珍妮特・诺伍德被任命为劳工统计局（Bureau of Labor Statistics）局长	J・诺伍德（Janet Norwood）
1980	埃贡・S・皮尔逊去世	E・S・皮尔逊
1981	耶日・奈曼去世	J・奈曼
1982	混沌理论（chaos theory）的现代表述	R・亚伯拉罕（R. Abraham）、C・肖（C. Shaw）
1983	表明个人概率局限性的研究	A・特韦尔斯基（A. Tversky）、D・卡内曼（D. Kahneman）
1985	哈拉尔德・克拉美去世	H・克拉美
1987	安德烈・尼古拉耶维奇・柯尔莫哥洛夫去世	A・N・柯尔莫哥洛夫
1987	将核回归（Kernel－based regression）应用到调焦照相机（“模糊系统”）	T・山川（T. Yamakawa）
1989	L・J・科恩对统计模型和方法的批评	L・J・科恩（L. J. Cohen）
1990	《观测数据的样条模型》（Spline Models for Observational Data）	G・沃赫拜（G. Wahba）
1992	鞅方法用于医学研究得到了充分发展	O・奥伦（O. Aalen）、E・安德森（E. Anderson）、R・吉尔（R. Gill）
1995	弗洛伦斯・南丁格尔・大卫去世	F・N・大卫
1997	将科克伦方法（Cochran's methods）扩展到序贯分析（sequential analysis）	C・詹尼森（C. Jennison）、B・W・特恩布尔（B. W. Turnbull）
1999	使 EM 演算法适用于有关奥伦－安德森－吉尔鞅模型的问题	R・A・比滕斯凯（R. A. Betensky）、J・C・林赛（J. C. Lindsey）、L・M・瑞安（L. M. Ryan）
2000	约翰・图基去世	J・图基